Bioinformatic analysis of biotechnologically important microbial communities

Submitted by Katy June Jones to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Biological Sciences in April 2018.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: .....

Abstract

Difficulties associated with the study of microbial communities, such as low proportions of cultivable species, have been addressed in recent years with the advent of a range of sequencing technologies and bioinformatic tools. This is enabling previously unexplored communities to be characterised and utilised in a range of biotechnology applications. In this thesis bioinformatic methods were applied to two datasets of biotechnological interest: microbial communities found living with the oil-producing alga *Botryococcus braunii* and microbial communities in acid mine drainage (AMD). *B. braunii* is of high interest to the biofuel industry due to its ability to produce high amounts of oils, in the form of hydrocarbons. However, a number of factors, including low growth rates, have prevented its cultivation on an industrial scale. Studies show *B. braunii* lives in a consortium with numerous bacteria which may influence its growth. This thesis reports both whole genome analysis and 16S rRNA gene sequence analysis to gain a greater understanding of the *B. braunii* bacterial consortium. Bacteria have been identified, some of which had not previously been documented as living with *B. braunii*, and evidence is presented for ways in which they may influence growth of the alga, including B-vitamin synthesis and secretion systems. AMD is a worldwide problem, polluting the environment and negatively impacting on human health. This by-product of the mining industry is a problem in the South West of England, where disused metalliferous mines are now a source of AMD. Bioremediation of AMD is an active area of research; sulphur-reducing bacteria and other bacteria which can remove toxic metals from AMD can be utilised for this purpose. Identifying bacteria and archaea that are able to thrive in AMD and which also have these bioremediation properties is therefore of great importance. Metagenomic sequencing has been carried out on the microbial community living in AMD sediment at the Wheal Maid tailings lagoon near Penryn in Cornwall. From these data have been identified a diverse range of bacteria and archaea present at both the sediment surface level and at depth, including microorganisms closely related to taxa reported from metalliferous mines on other continents. Evidence has been found of sulphur-reducing bacteria and of pathways for various other bioremediation-linked processes.

Acknowledgements

I would like to express my gratitude to my supervisors Steve Aves and David Studholme for all of your help and support throughput my PhD. Steve, our regular meetings were invaluable in keeping me motivated and I have appreciated every piece of advice you offered me over the years. Thank you for investing so much time and patience in me! I would also like to thank John Love, Mark van der Giezen and Chris Bryan for allowing me to collaborate with them on various projects and Karen Moore for helping me plan my sequencing projects. I would also like to acknowledge the BBSRC, who funded this research.

Thank you to all members of the Mezzanine lab (past and present) for all the help and laughter, especially Christine Sambles for always being there to answer my bioinformatic questions, Richard Tennant for your endless help when things went wrong and Ann Power for all the motivational tea breaks. You all played a major part in picking me up when I felt like giving up. Thank you.

Thank you to my husband Garan and my two wonderful children, Molly and Morgan, for putting up with being neglected whilst I wrote my thesis! Molly and Morgan, I hope seeing your mummy achieve her PhD will show you that you can achieve anything you want to in life, whatever that may be.

Finally, I owe a world of thanks to my mum. I would never have been able to complete any of my time at university, from my undergraduate through to my PhD, without your constant support and provision of free childcare! I don't thank you often enough, but I hope you know how much I love and appreciate you.

TABLE OF CONTENTS

List of figures

List of tables

Abbreviations

AMD Acid Mine DrainageANI Average Nucleotide Identity
BAC Bacterial Artificial Chromosome
BLAST Basic Local Alignment Search Tool
GFF General Feature File
LCA Lowest Common Ancestor
LCM Lignocellulosic Materials
MLST Multilocus Sequence Typing
NGS Next Generation Sequencing
OLC. Overlap-Layout-Consensus
OUT Operational Taxonomic Unit
PCR Polymerase Chain Reaction
QC Quality Control
RPKM Reads Per Kilobase per Million
SRB Sulphate Reducing Bacteria
WGS Whole Genome Sequencing

Chapter One: Introduction

1.1 The unexplored world of microbial communities

Microbial life is found living in every habitat on earth. Conditions once thought too hostile to possibly host life are now known to be home to a range of microbial communities. Microbial life impacts upon the entire biosphere, acting as the driving force behind every basic ecosystem process on earth (Whitman *et al.*, 1998). But despite their crucial role in all aspects of life, large numbers of microbes remain uncharacterised and an understanding of the microbial community structure of a large range of habitats is still lacking, mainly due to the high biodiversity of most microbial ecosystems and the inability to cultivate the majority of microbes from the environment (Rappe and Giovannoni, 2003; Eloe-Fadrosh *et al*., 2016).

1.1.1 Methods of studying microbial communities

Traditionally the first step in the study of microbial life has been to culture microbes, allowing them to multiply in a culture medium within a laboratory setting. However, the realisation that a vast number of bacteria within a given environmental sample cannot be cultured came about in 1985, with Staley and Konopka (1985) coining the term 'the great plate count anomaly' to describe the large difference between the number of cells observed through a microscope and the number forming colonies on an agar medium. It has since been determined that only around 1 % of all bacteria are able to be easily cultivated *in vitro* (Vartoukian *et al.*, 2010). The most widely accepted explanation for the plate count anomaly is that microbes require particular conditions in which to grow, such as certain levels of nutrients, oxygen, temperature or pH and do not tolerate deviations from the levels found in their natural environment. Many microbes also require more oligotrophic conditions than those provided by typically used media (Vartoukian *et al.*, 2010). A 2008 study by Nichols *et al*. investigated the possibility that certain microorganisms will not grow in culture as they are reliant upon other microorganisms present in their native environment. Nichols *et al*. found that by placing microorganisms into a diffusion chamber which was returned to the native environment during incubation, they were able to culture and isolate microorganisms which previously had refused to grow *in vitro*. Furthermore, these microorganisms were able to grow on

artificial media if accompanied by 'helper strains' from their environment, suggesting growth promoting signals from other members of the microbial population were enabling this to happen. Although there is now an increasing number of culture-independent methods for the analysis of bacteria (which will be discussed in subsequent sections), culturing has many advantages, notably including the isolation of individual species allowing for physiological properties and virulence potential to be fully assessed (Vartoukian *et al.*, 2010). Therefore, new methods to culture the currently uncultivable have been widely explored following the revelation of Staley and Konopka (1985), with new types of media being developed in order to target a wider range of organisms. However, despite these efforts, the majority of microbial life remains currently uncultivable (Sherpa *et al*., 2015).

The use of 16S rRNA for taxonomic classification

As details of the great plate count anomaly became more widely understood, methods for studying microbial communities that did not rely on culturing were further explored. In 1977 Woese and Fox published their pioneering work on 16S rRNA that was to revolutionise the scientific world's understanding of microbial life. Woese and Fox demonstrated that cellular life can be divided into three domains, as well as defining 11 separate phyla within which the domain Bacteria can be further classified. Woese achieved this through the phylogenetic analysis of 16S rRNA from cultivated micro-organisms, with sequence differences being represented as divergences in a phylogenetic tree. Although Woese's analysis was cultivation-dependent, methods for cultivation-independent 16S rRNA gene sequencing were soon in development and by the early 1990s studies analysing bulk 16S rRNA gene sequences from environmental samples were beginning to emerge (Lane *et al*., 1985; Giovannoni *et al*., 1990; Schmidt *et al*., 1991). These early studies often made use of PCR amplification of 16S rRNA genes and Sanger sequencing. Today, 16S rRNA is the most sampled gene in prokaryotes (Richards and Bass, 2005; Glockner *et al*., 2017). Its popularity in the study of microbial populations is due to a number of factors: 16S rRNA is present in all prokaryotes, it is highly conserved between species and it contains conserved and variable regions which can both be utilised for primer design and taxonomic classification. At around 1500 bp, 16S rRNA is also a good size for bioinformatics analysis

(Janda and Abbott, 2007). However, despite its range of advantages and significant contribution to the current number of recognised bacterial taxa, the use of 16S rRNA sequence analysis in phylogenetics is not without its drawbacks. Although 16S rRNA sequences have been described as the gold standard of bacterial classification they often have low resolution at the species level especially between species of the same genus. Species of bacteria which are highly distinguishable biochemically can have 16S rRNA sequence identity of over 95% (Janda and Abbott, 2007; Fox *et al*., 1992) The quality of 16S databases is also of concern; misidentified strains contribute to the 'pollution' of databases, with well-studied genera being especially affected by this (Mignard & Flandrois, 2006). Mixed cultures may also form chimeric molecules which can lead to chimeric sequences being deposited into databases, essentially describing non-existent species (Hugenholtz & Huber, 2003; Mignard & Flandrois, 2006). The use of other conserved 'housekeeping' genes in taxonomic studies, known as multilocus sequence typing (MLST), may offer more accurate identification at and below the genus level than through 16S rRNA sequence analysis alone. Martens *et al*. (2008) discussed the advantages of using this approach in a study in which they carried out phylogenetic analysis on ten housekeeping genes (*atpD*, *dnaK*, *gap*, *glnA*, *gltA*, *gyrB*, *pnp*, *recA*, *rpoB* and *thrC*) from 34 representatives of the genus *Ensifer*. Comparisons with 16S rRNA gene sequences demonstrated a much higher discrimination potential as well as clearer species boundaries for all ten housekeeping genes. The benefits of using these protein-coding housekeeping genes over those encoding 16S rRNA have been further explored, with the discriminatory power of the *gyrB* and *rpoB* genes shown to be better than 16S rRNA, even in species which had almost identical 16S rRNA gene sequences (Carrasco *et al.*, 2013). The discriminatory power of housekeeping genes can be measured on the Hunter-Gaston discriminatory index; this index was created in 1988, based on the formula of Simpson's index of diversity, to provide an average probability that a typing system will assign different types to two unrelated strains randomly sampled in the microbial population of a given taxon (Hunter and Gaston, 1988). This can be a useful tool when selecting housekeeping genes.

Metagenomics and sequencing methods

The term 'metagenomics' was first used in a study by Handelsman *et al.*, published in 1998. Handelsman *et al*. aimed to analyse the genomes of microbial life found in soil, in order to better understand those organisms which could not be traditionally cultured. In order to achieve this objective, Handelsmans *et al*. described their methods for cloning the soil metagenome; a soil sample was obtained and bacterial DNA fragments were isolated, cut with restriction enzymes, cloned into bacterial artificial chromosomes (BACs) and transformed into *Escherichia coli* before being screened for novel enzymes. However, despite Handelsman's study coining the phrase metagenomics, a more widely accepted meaning of the term was defined by two studies in 2004 (Venter *et al*., 2004; Tyson *et al*., 2004). These two studies, used whole genome shotgun (WGS) sequencing to analyse environmental samples. WGS had previously been applied to the study of whole genomes of individual organisms. The process of WGS sequencing starts with the random, mechanical or chemical shearing of the genome into smaller fragments which are then cloned in BACS before being sequenced (using Sanger sequencing in these early studies). Whilst Tyson *et al.* utilised these methods to investigate a small range of microorganisms found in an extreme acidic environment, Venter *et al*. used them to study a large range of microorganisms found in a sample of seawater. These studies demonstrated the ability of WGS sequencing methods to be applied to two different microbial communities and set the stage for the wide range of studies on microbial communities using WGS sequencing, which soon followed.

The advent of Next Generation Sequencing (NGS) methods, in 2005, revolutionised genome projects and led to cheaper, faster DNA sequencing (Margulies *et al*., 2005; Schuster, 2007). Although numerous NGS platforms were developed, including Illumina, Roche 454 Pyrosequencing and ABI SOLID (Rodrigue, 2010), the general principles behind them remain the same. These high-throughput methods eliminate the need for the bacterial cloning step used in traditional Sanger sequencing, instead they amplify DNA molecules and parallelise the sequencing process, meaning huge numbers of sequences can be generated concurrently (Bubner, 2008). However, this faster, cheaper sequencing came at the cost of read length; whilst Sanger sequencing typically

produced sequence reads of around 400-900 base pairs, early NGS methods generally produced much shorter reads (Rodrigue, 2010). However, in recent years, advances in sequencing methods have seen improvements to the read lengths of NGS-generated sequences. Illumina has been the dominant NGS technology in terms of market share and technological developments (Goodwin *et al*., 2016). Current Illumina technology includes bench-top sequencer the MiSeq, which can produce read lengths of 300 bp, output 25 million reads per run and has a run time of just 4-55 hours; the MiSeq is frequently used for 16S microbial community studies. Illumina's production scale sequencers include the HiSeq series, which produces shorter read lengths of 150bp and can take up to 6 days to run, but has an output of up to 5 billion reads per run; the HiSeq is a good choice for shotgun metagenomic projects (Illumina, 2018).

Next generation sequencing has revolutionised the field of genomics. However, there is still room for new technologies; so-called "third-generation" sequencers which take a different approach than NGS. Oxford Nanopore and Pacific BioSciences (PacBio) are the two main companies that have brought further advancements to this area. Both Oxford Nanopore and PacBio sequencing technologies utilise single-molecule sequencing and produce read lengths much longer than NGS methods with an average of 12 kb for PacBio and over 300 kb for Oxford Nanopore (Giordano *et al*., 2017; Jain *et al*., 2016). The 'MinIon', by Oxford Nanopore is a portable sequencing device which plugs directly into a USB port and has enabled the rapid real-time sequencing of clinical and environmental samples and was used as a tool for on-site monitoring of the 2013-2016 West Africa Ebola virus epidemic (Quick *et al*., 2016). The PacBio's Single Molecule Real Time sequencing (SMRT) technology is a production scale sequencer which, like the MinION from Oxford Nanopore Technologies, is offering great advancements in genomic studies due to the long-read lengths it can produce (Rhodes & Au, 2015). However, the long reads generated by third generation sequencers are more error-prone than NGS methods and analysis can incur high computational costs (Chin *et al*., 2016; Xiao *et al*., 2017).

Advances in next generation sequencing methods have led to significant breakthroughs in the study of microbial ecosystems and a greater

understanding of the dynamics of microbial communities (Oulas *et al*., 2015). Enabling faster, cheaper sequencing, resulting in millions of reads and offering insights into the world of currently uncultivable microorganisms, NGS metagenomic studies have revealed the huge diversity present within many microbial ecosystems (Culligan *et al*., 2016). Microbial communities have been studied from a diverse range of environments across the globe and the microbiomes of large numbers of animals have been characterised and to date there are 51,828 metagenomes available (Metagenomics RAST server; http://metagenomics.anl.gov; publicly available metagenomes as of February 2018). One area in which the use of metagenomics has further advanced knowledge of microbial communities is that of extreme environments. Extremophilic organisms are found living in environments that experience extremes of temperature, pH, salinity, pressure, radiation levels or toxins. Organisms in these environments are notoriously difficult to culture due to their need for niche environmental conditions and metagenomics is therefore an especially useful tool in this field (Cowan *et al*., 2015).

Advances in sequencing methods have enabled individual whole genome sequencing to be faster, cheaper and more accurate, with public repositories (as of February 2018) containing over 130000 entries for prokaryotic genomes (https://www.ncbi.nlm.nih.gov/genome/browse/#!/prokaryotes/). The majority of these genomes are from axenic cultures and there is a bias towards medically important organisms (Parks *et al*., 2017; Kyrpides *et al*., 2013), therefore, in recent years efforts have been made to obtain reference genomes from more diverse phylogenies, with the Genomic Encyclopedia of Bacteria and Archaea (GEBA) initiative recently sequencing 1003 genomes in order to expand the current genome database by including prokaryotes with a broad phylogenetic and physiological diversity (Mukherjee *et al*., 2017). However, although the value of whole genome sequencing from individual organisms is high, the expansion of current genome collections still only includes those that can be cultivated (Dini-Andreote *et al*., 2012). A significant breakthrough in the problem of the diversity of whole genomes available has been provided by recent advances in metagenomics (Parks *et al*., 2017). Reconstructing near-complete whole genomes from metagenomic datasets is now a possibility and whereas this was previously only possible from metagenomic samples with low species

diversity this can now be achieved with high diversity samples, due to improved throughput and computational techniques. Recently, the first large scale project to attempt this recovered nearly 8000 draft-quality genomes from over 1500 metagenomes, with the authors claiming this had substantially expanded the tree of life (Parks *et al*., 2017). The retrieval of whole genomes from metagenomes is important progress in exploring currently uncultivable organisms, however it is still a developing method, with the requirement of very high coverage of communities, and the majority of genomes retrieved are only of 'near-complete' standard. Therefore, culture-dependent methods of retrieving whole genomes are likely to still be used alongside metagenomics while further advances are made.

Metagenomics is a highly valuable tool in the analysis of the structure of microbial communities. Genomic information allows the key question of "Who is out there and what are they doing?" to be addressed (Escobar-Zepeda *et al*., 2015). However, despite the significant contribution metagenomics has made to the study of microbial ecology, it has its limitations and although it can be used to determine which genes are present, it offers no insights into which metabolic functions are active or suppressed within a community at a given point in time. This is an important consideration when studying topics such as a community's response to changes in its external environment (Gilbert and Hughes, 2011). Metatranscriptomics was developed to enable a better understanding of this area, allowing for researchers to record the ribosomal and messenger RNA being actively transcribed at a set point of time or over a period of time (Gilbert and Hughes, 2011). Like metagenomics, metatranscriptomics can be applied to mixed communities of microbes; however, whereas metagenomics involves the isolation of DNA from a sample, metatranscriptomics isolates the messenger RNA (mRNA), non-coding RNAs and small RNAs present. Early studies would then analyse the mRNA through microarray technology, or cDNA clone libraries would be derived from the mRNA. However, both these methods have limitations, with bias introduced in cloning and insensitivity to very high or low expressed sequences in microarrays (Simon and Daniel, 2010). These limitations can now be overcome by using next generation sequencing methods; this involves the isolation of RNA which is then reverse-trancribed to cDNA and sequenced using a high-throughput method such as Illumina. These

reads are then able to be assembled against reference genomes/transcripts or can be assembled *de novo.* This makes the sequencing of RNA (RNA-seq) a highly useful tool when studying microbial transcripts of unknown organisms which do not already have a fully sequenced genome (Wang *et al.*, 2009).

Advances in the study of microbial communities and genomics have led to the production of increasingly large sequencing datasets. While the cost of sequencing is falling, the computational requirements to deal with the huge amounts of data often produced is increasing and must be considered when planning metagenomic studies (Su & Ning, 2012). There are now numerous pipelines and software packages available for the analysis of whole-genome or metagenomic sequence data which can be tailored depending on the questions that need answering and the type of dataset being analysed. Key stages in the analysis of sequencing datasets include quality control, sequence assembly, taxonomic classification and annotation. These are discussed further throughout this thesis.

1.2 Microbial communities and biotechnology

The study of microbial life is important in fully understanding the ecology of a range of habitats across the world. However, microbes can also be utilised in numerous biotechnology applications in the areas of medicine, environment, agriculture and industry. These have typically involved the use of single strains such as *Escherichia coli,* which is used in the synthesis of medicinal products including insulin, interferon and human growth hormone (Goeddel *et al*., 1979; Ikehara *et al*., 1984; Yelverton *et al*., 1981), or *Agrobacterium,* used in the genetic engineering of a range of plants (Gelvin, 2003). However, there are also examples of the use of microbial communities in biotechnology; bioleaching involves the use of iron-oxidising bacterial communities to extract metal from mines and bioremediation often utilises mixed microbial communities to aid in the clean-up of environmental pollutants (Leal *et al*., 2017). The rising interest in biofuels in recent years has frequently involved research into the use of microbial communities; optimum communities for the production of biogas or the

breakdown of lignocellulosic biomass have been explored (Rubin, 2008; Stolze *et al.*, 2015) and studies conducted on the effects of microbial communities on the growth and yield of algal biofuels (Wang *et al*., 2016). Next generation sequencing and metagenomics have also allowed the mining of microbial communities for enzymes that may prove useful in biofuel synthesis (Xing *et al*., 2012).

This thesis focuses on two areas of biotechnology and bacterial communities: the potential use of microbial communities to increase production of algal biofuels and the impact of microbial communities on the production and bioremediation of acid mine drainage, which will be looked at in more detail in sections 1.2.1 and 1.2.2.

1.2.1 The impact of microbes on algal biofuel production

With global energy demands ever increasing and concerns over global warming growing, biofuels were initially hailed as a green, sustainable alternative to fossil fuel (Asveld, 2016). First generation biofuels, including biodiesel, biogas and ethanol have become widely used throughout the world and the production process used is now considered established technology (Naik *et al*., 2010). However, their use has not been without controversy; the food versus fuel debate branded the use of food crops for biomass unethical and it is claimed the land-use changes associated with large scale production of biofuel crops has had a detrimental impact on the environment as well as generating increased carbon emissions and a large water-footprint (Searchinger *et al*., 2008). Second generation biofuels aimed to mitigate some of these issues through the use of non-food crops and non-edible parts of food crops such as stems and husks, known as lignocellulosic material (LCM). However, technical difficulties with the processing of LCM into liquid fuels currently mean second generation biofuels are not cost effective (Bhatia *et al*., 2017).

Previously labelled “second generation”, algal biofuels have now been reclassified as third generation biofuels, due to the realisation that they can produce higher yields and use fewer resources than biofuels from feedstock; microalgae biomass can be grown in photobioreactors, open ponds or closed-loop systems (Figure 1.1) using fresh, saline, brackish or even waste water

Figure 1.1 Algal biofuels being grown in, from top to bottom,: Open ponds, photobioreactor and a closed loop system.

Photo credits: Karen Fehrenbacher

meaning it is not in competition with arable land (Behera *et al*., 2015). Additionally, algal biomass can be used to produce a diverse range of fuels including biodiesel, butanol, ethanol, petrol (gasoline) and jet fuel. However, algal biofuels also have their limitations, with high costs associated with infrastructure, biomass harvesting and lipid extraction (Kim *et al*., 2014). Critics of algal biofuel have raised concerns that the resources, including energy and nutrients, required to meet just 5 % of the transportation fuel needs in the U.S.A are highly impractical (Waltz, 2013). It is therefore imperative to identify strains of algae which produce the highest volumes of oil possible, as well as identifying optimal growing conditions in order to make the sustainable production of algae biofuel a real possibility.

The name algae encompasses a diverse polyphyletic group of photosynthetic organisms, ranging from unicellular microalgae to giant kelp (brown algae) which can grow up to 50 metres. There are 15 divisions and 54 classes of algae, with estimates of species numbers ranging from thirty thousand to over one million (Guiry, 2012). Of interest to the biofuel industry are the microalgae, with the green algae (consisting of the divisions Chlorophyta and Charophyta) most frequently used for this purpose (Scott 2010). Along with factors such as ease of harvest, resistance to contaminants and tolerance to a wide range of culture parameters, a key component of microalgae in respect to its use as a biofuel is lipids, with its usefulness to the biofuel industry dependent on both growth rate and amount of lipids produced (Griffiths & Harrison, 2009). Table 1.1 shows the lipid content of a range of microalgae. The lipids accumulated by microalgae include both nonpolar storage lipids, such as triacylglyceroles (TAGs) and hydrocarbons as well as polar lipids, such as glycolipids, phospholipids and sterols which are found in cellular membranes (Georgianna & Mayfield, 2012). The green alga *Botryococcus braunii* is of high interest to the biofuel industry due to its ability to accumulate large quantities of lipids in the form of hydrocarbons (up to 75 % of its dried weight). *B. braunii* is discussed in more detail in chapter two.

Due to the potential microalgae have as a good source of renewable energy, extensive research has been carried out into ways in which growth and production of oils can be increased to make algal biofuels more cost effective. A

Table 1.1 A range of microalgae and their lipid content (Adapted from Ahmed *et al*., 2017)

Microalgae	Division	Oil content (% dry weight biomass)
Ankistrodesmus sp.	Chlorophyta	24-31
Botryococcus braunii	Chlorophyta	25-75
Chaetoceros muelleri	Heterokonta	33
Chlamydomonas reinhardtii	Chlorophyta	21
Chlorella spp.	Chlorophyta	10-48
Crypthecodinium cohnii	Dinoflagellata	20-51
Dunaliella tertiolecta	Chlorophyta	16-71
Ellipsoidion sp.	Ochrophyta	27
Euglena gracilis	Euglenozoa	14-20
Haematococcus pluvialis	Chlorophyta	25
Nannochloris spp.	Chlorophyta	20-56
Neochloris oleoabundans	Chlorophyta	29-65
Nitzschia spp	Heterokonta	45-47
Phaeodactylum tricornutum	Heterokonta	18-57
Prymnesium parvum	Haptophyta	22-39
Scenedesmus obliquus	Chlorophyta	11-55
Schizochytrium sp.	Heterokonta	50-77
Thalassiosira pseudonana	Heterokonta	20

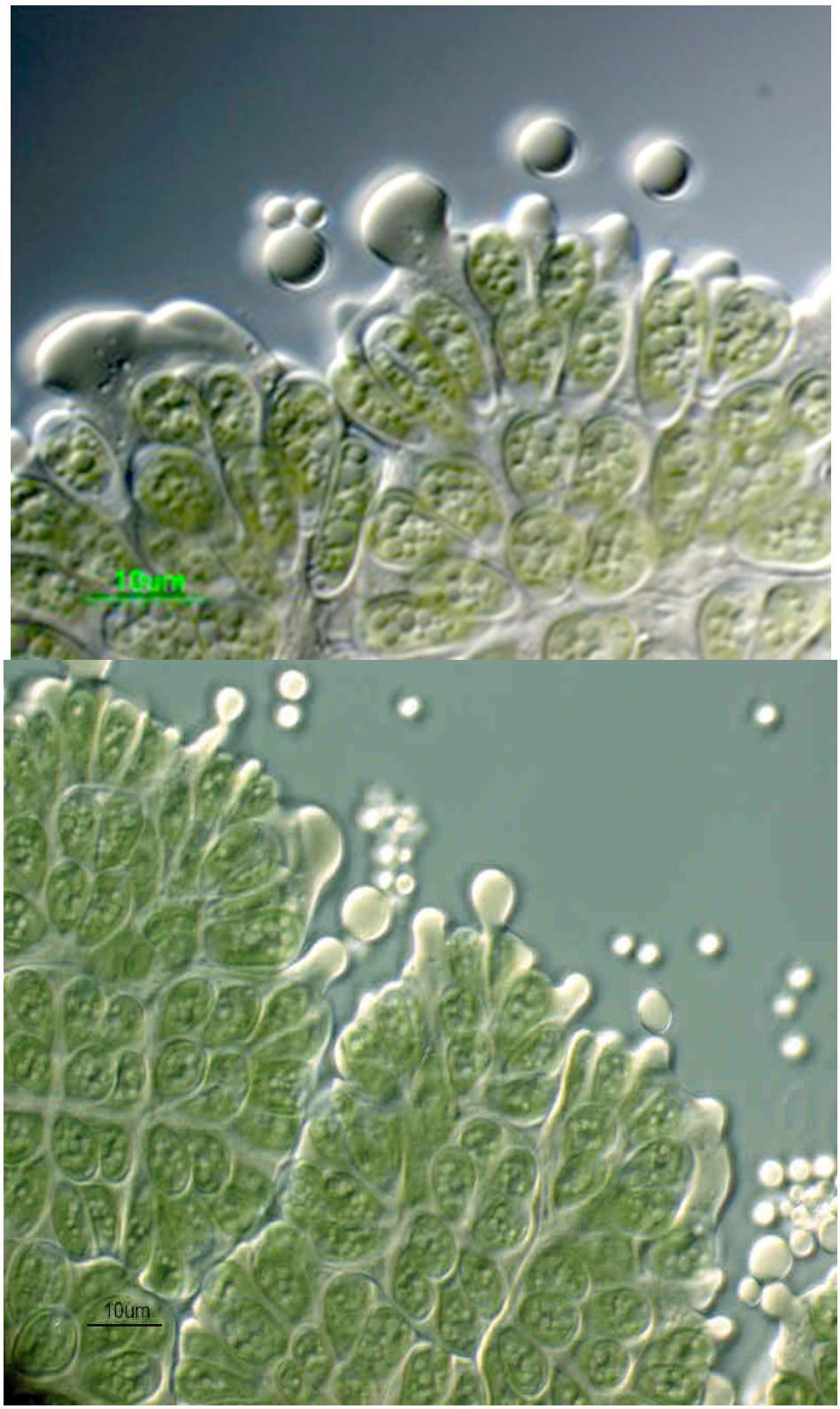


Figure 1.2 Microscopy images of *B. braunii*. Algal cells are held together by an extracellular matrix composed of hydrocarbons, hydrocarbons can also be seen exuding from the matrix and as small drops within the alga cells. Photo credit: Top, Tim Devarenne (Texas A&M). Bottom, alchetron.com/Botryococcus-braunii

full understanding of microbial communities associated with microalgae is now of greater interest to the algal biofuel industry due to studies that have shown algae-bacterial symbiosis may have a positive effect on algal biomass production; as well as increasing growth rates bacterial interactions may assist in algal flocculation (Fuentes *et al*., 2016). One of the costliest steps of producing algal biofuel is the harvesting of biomass; typically, this involves attempting to separate a low mass of microalgae from a large volume of water (Uduman *et al*., 2010). This dewatering process is expensive, with common methods including centrifugation, filtration, electrocoagulation, forward osmosis and flocculation (Gerado *et al*., 2015; Lee *et al*., 2013). One of the cheapest methods for partial dewatering is flocculation followed by gravity sedimentation of flocs (Chatsungnoen & Chisti, 2016). Flocculants used in this process include inorganic metal salts such as ferric sulphate, ferric chloride, aluminium sulphate and aluminium chloride as well as organic materials such as biopolymers. However, inorganic methods can be toxic and negatively affect the viability of algal cells, whilst organic methods are not cost effective (Ummalyma *et al*., 2017). Bioflocculation involves the use of microorganisms such as bacteria, fungi or other algae to induce flocculation and has recently gained interest as a method which may help to alleviate some of the previously mentioned harvesting problems (Alam *et al*., 2016). Many bacteria are likely to have the ability to increase flocculation in microalgae those studied so far include: *Paenibacillus* sp., *Solibacilus silvestris, Flavobacterium* sp., *Terrimonas* sp. and *Sphingobacterium* spp (Wan *et al*., 2013; Lee *et al*., 2013). The mechanisms that drive bacterial-induced microalgae flocculation are not fully understood, but are likely to involve both direct interactions between the algae and bacteria, where physical contact between the two causes aggregation as well as indirect interactions, in which the release of bacterial compounds causes aggregation (Alam *et al*., 2016). Growth rates are another important factor in mass producing algal biomass and studies have shown that some bacteria positively affect this, with different bacterial communities found in consortia with a wide range of microalgae (Ramanan *et al*., 2016).

The relationships that have been observed occurring between bacteria and microalgae encompass all types of symbiosis: mutualism, commensalism and parasitism (Ramanan *et al*., 2016). Bacteria may be found as a biofilm on the

algal surface as well as in the region immediately surrounding algal-cells the phycosphere – or in looser association within the water column (Dang & Lovell, 2016). Gaining greater insight into these interactions and the bacteria involved may lead to the controlled utilisation of microalgae-bacteria consortia to increase biomass for the biofuel industry (Fuentes *et al*., 2016).

In mutualistic algal-bacterial relationships both partners benefit from each other. A typical example of this is the exchange of vitamins produced by bacteria and fixed carbon produced by algae. Many algae have been found to be auxotrophic for various combinations of B-vitamins: vitamin B12 (cobalamin), vitamin B1 (thiamine), and vitamin B7 (biotin). A 2006 study which surveyed 306 species found more than half required cobalamin, 22 % required thiamine and 5 % required biotin (Croft *et al*., 2006). The concentration of these vitamins in the natural environment are usually very low, and it was determined that some or all of these vitamins are most likely to be provided by bacteria living alongside algae. This has been demonstrated most frequently when looking at vitamin B12, as only prokaryotes are able to synthesise this vitamin, meaning all the B12 found in algae must originally have been produced by bacteria (Croft *et al*., 2005). Bacteria are able to survive in media with no carbon source when in an algal-bacterial consortium, however, the equilibrium between algal and bacterial cells can be disrupted if an additional carbon or B12 source is added, and neither can usually survive when the two are separated demonstrating the mutualistic nature of the relationship (Kazamia *et al*., 2012).

Symbiotic relationships involving plant growth promoting bacteria (PGPB) are well documented in plants; nitrogen fixation by rhizobia in exchange for organic acids as a carbon and energy source is one such process that occurs in the rhizosphere and is a clear example of mutualism (Denison & Kiers, 2004). The effects of PGPB on microalgae have also been studied, though less extensively than plants, with the phycosphere of algae being analogous with the rhizosphere of a plant for their ability to host colonies of bacteria (Kim *et al*., 2014). As in plants, the fixation of nitrogen is a key process in the relationship between bacteria and algae, with bacteria carrying out this function while benefitting from the fixed organic carbon supplied by algae (Cho *et al*., 2015). A 2014 study found *Rhizobium* to be the dominant bacterial genus in the

phycosphere of strains of the green algae *Chlamydomonas reinhardtii*, *Chlorella vulgaris*, *Scenedesmus* sp., and *Botryococcus braunii*. These algae were then co-cultured with *Rhizobium* sp. which was found to enhance algal growth by 72 %. Additionally, the growth of *Rhizobium* sp. was significantly increased when in the presence of all of these algae (Kim *et al*., 2014). The nature of the mutualistic relationship that was occurring was not fully determined, but *Rhizobium* are nitrogen-fixing bacteria found in the rhizosphere of plants and this was likely playing a role in its ability to enhance algal growth. Other PGPB that have been shown to have a positive effect on green algal growth include *Bacillus pumilus* and *Azospirillum brasilense* which were shown to increase growth through nitrogen fixation and the provision of phytohormones respectively (Hernandez *et al*., 2009; Gonzalez & Bashan, 2000).

In algal-bacterial commensalism only one of the partners benefits whilst the other remains unaffected by the interaction taking place. One such example is that of a bacterium which is not providing vitamins or other nutrients to an alga but is living in a commensal relationship whereby it is benefiting from the carbon source the alga supplies (Bratbak & Thingstad, 1985). It has been suggested that shifts between mutualism and parasitism may occur via commensalism and that these shifts in the nature of the symbiotic relationship are driven by environmental factors (Fuentes *et al.,* 2016). Studies have shown that the balance of algal-bacterial communities changes depending on nutrient availability within the environment. Nitrogen and phosphorous are two nutrients that both algae and bacteria usually need to acquire from their environment; when these nutrients are limited they become competitors (Danger *et al*., 2007). Bacteria are better competitors for phosphorous than algae and drops in phosphorous levels have been linked to previously steady communities becoming outnumbered by bacteria (Grover, 2000; Gurung *et al*., 1999). As nutrient levels in the environment drop, algae become stressed and release higher levels of organic carbon which is then utilised by the bacteria, increasing bacterial numbers and further increasing the competition for nutrients (Bratbak & Thingstad, 1985).

As well as negatively affecting algal growth through nutrient competition, some bacteria also have algicidal properties. Although bacterial inhibition of biofuel-

producing microalgae has not been widely studied, algicidal bacteria are well known for their interactions with naturally occurring microalgal blooms (Wang *et al*., 2016). The majority of algicidal bacteria belong to the phyla Bacteroidetes or Proteobacteria and include the genera *Alteromonas*, *Pseudomonas* and *Pseudoaltermonas (Meyer et al., 2017).* Lysis of algal cells appears to be the most common mechanism by which algicidal bacteria kill algae; this can occur either through direct contact of the bacteria with the algae or through indirect contact whereby extracellular algicidal compounds are released by the bacteria (Mayali & Azam, 2004). The nature of the relationship between bacteria with algicidal properties and microalgae is likely to alter depending on both environmental factors and algal growth. The two may happily exist in mutualistic or commensal symbiosis for much of the time, however when algal blooms develop and algal exudates increase, the number of algicidal bacteria also increases, lysing algal cells which in turn provide further nutrients for bacterial growth thus accelerating the decline of the bloom. As algal blooms terminate and microalgal exudates fall, the abundance of bacteria and alga return to their previous stable levels (Doucette *et al*., 1999). Although the majority of studies have looked at algicidal bacteria in relation to methods of dealing with harmful microalgae and cyanobacterial blooms, it has also been suggested that algicidal bacteria could be utilised in the extraction of lipids from algal biofuels by using them to assist in the degradation of the algal cell wall, lessening the need for more expensive extraction methods currently in use (Lenneman *et al*., 2014). Two strains of bacteria that were identified as potentially useful for this purpose were *Pseudomonas pseudoalcaligenes* AD6 and *Aeromonas hydrophila* AD9 due to their cell degrading properties (Lenneman *et al*., 2014).

1.2.2 Microbial communities and acid mine drainage (AMD)

Acid mine drainage (AMD) is the biggest environmental concern associated with coal and mineral mining, having a detrimental effect on human health and negatively impacting on the ecology of thousands of working and disused mine sites throughout the world (Sheoran & Sheoran, 2006). Figure 1.3 shows the dramatic impact AMD has on the landscape. The production of AMD occurs when sulphide-bearing minerals, usually iron pyrite, undergo a serious of chemical reactions having been exposed to weathering through the mining

Figure 1.3 The impact of Acid Mine Drainage on the landscape; the presence of iron and other metals creates dramatic colours in, clockwise from top left: river run off from the Richmond Mine (Iron Mountain, USA), Wheal Maid tailings lagoon (Cornwall, UK), AMD tailings in Gauteng (S. Africa) and The Tinto River (SW Spain).
Photo credits: @carnkie, Mike Thomas, environment.co.za and sakuraamazingplace.blogspot.co.uk

process. Firstly, the pyrite (FeS_2) is oxidised, forming dissolved iron, sulphate and hydrogen ions:

$$2\ FeS_2 + 7\ O_2 + 2\ H_2O = 2\ Fe^{2+} + 4\ SO_4^{2-} + 4\ H^+ \quad (1)$$

If the environment is sufficiently oxidising the resulting ferrous iron then further oxidises to ferric iron.

$$4\ Fe^{2+} + O_2 + 4\ H^+ = 4\ Fe^{3+} + 2\ H_2O \quad (2)$$

The resulting Fe^{3+} further oxidises the pyrite producing ferrous iron , sulphate and hydrogen ions:

$$FeS_2 + 14\ Fe^{3+} + 8\ H_2O = 15\ Fe^{2+} + 2\ SO_4^{2-} + 16\ H^+ \quad (3)$$

In this third reaction iron (Fe^{3+}) is the pyrite oxidising agent, which has a higher rate of abiotic oxidation than O_2 and water, and the production of hydrogen ions at this stage results in a significant drop in pH causing the system to become highly acidic. If the pH remains above 3.5 ferric iron precipitates, forming $Fe(OH)_3$ also known as “yellow boy” due to its yellow-orange colour and more hydrogen ions are also produced:

$$4\ Fe^{3+} + 12\ H_2O = 4\ Fe(OH)_3 + 12\ H^+ \quad (4)$$

An overall summary reaction for the production of AMD is:

$$4\ FeS_2 + 15\ O_2 + 14\ H_2O = 4\ Fe(OH)_3 + SO_4^{2-} + 16\ H^+ + \text{heat} \quad (5)$$

The acidity of the mine drainage also results in the mobilisation of metals previously contained within the mine environment; typically, this includes zinc, cadmium, copper, arsenic, manganese, lead and nickel as well as other metals and metalloids depending on the composition of rocks and minerals in the locality of the AMD. The toxicity of many of the metals present within AMD poses a threat to the local environment and human health, although the specific risks vary from site to site (Akcil & Koldas, 2006). Primary factors which influence the rate of acid generation are: pH, temperature, oxygen levels,

saturation levels, surface area of exposed metal sulphide and microbial activity. Despite the harsh conditions of low pH, high temperatures and presence of toxic metals, a diverse range of microbial life has been found living in AMD (Baker & Banfield, 2003) some of which facilitate the oxidation of iron (reaction 2) leading to the acceleration of pyritic oxidation (reaction 3) by up to 10^6 times, increasing the overall rate of AMD production significantly (Mielke *et al*., 2003). Bacteria which have been shown to oxidise iron, and play a key role in AMD production include: *Acidithiobacillus ferrooxidans* (AKA *Thiobacillus ferrooxidans*), *Leptospirillum ferrooxidans*, *Acidithiobacillus thiooxidans* and *Sulfobacillus thermosulfidooxidans*. Iron-oxidizing acidophilic archaea have also been found in AMD including *Ferroplasma acidiphilum* and *Thermoplasma acidophilum* with more species expected to be discovered (Chen *et al*., 2016; Mendez-Garcia *et al*., 2015; Tyson *et al*., 2004). In some circumstances iron-oxidising microbes are actively encouraged to facilitate in the mining process (Zammit *et al*., 2012); bioleaching utilises bacteria such as *Acidithiobacillus* and *Leptospirillum* in the extraction of metal from ores (usually in the mining of copper, nickel, cobalt, zinc and uranium) through the oxidation of sulphur and iron, however, although this is an efficient way of obtaining metals from low-grade ores which may be more difficult to mine using traditional methods, bioleaching still results in the formation of AMD.

The treatment for AMD varies depending on the site but generally falls into one of two categories: active or passive. Active treatment requires the input of power to pump water to treatment plants where neutralising agents such as limestone, hydrated lime, caustic soda or ammonia are mixed with AMD, raising the pH and causing metals to precipitate out of solution. Passive treatments use geochemical or biological processes (or a combination of the two) to neutralise AMD; geochemical methods include the use of limestone channels/drains, limestone leach beds, diversion wells, and limestone sand, while biological methods include the use of anaerobic/aerobic wetlands, vertical flow wetlands, and sulphur-reducing bioreactors (Skousen *et al*., 2017). The use of microorganisms to facilitate in the neutralisation of AMD and the removal of metals is a key component in passive treatments and is discussed further below. Both active and passive treatments have benefits and drawbacks; active systems can treat high and fluctuating flow rates as well as any pH range but

have higher running costs and require a power source (Taylor *et al*., 2005), while passive treatments have lower running costs and do not require power but require a larger treatment area and are typically more effective at treating low flow rates of AMD within a set pH range (Skousen *et al*., 2017).

Microbial activity occurs in all biological passive treatments of AMD with varying degrees of impact on remediation. Aerobic wetlands are used for the treatment of AMD that has been pH neutralised, often through another biological or geochemical passive process, and enables metal removal through the oxidation of Fe_2 to Fe_3 which then co-precipitates with other metals present; in this system the oxidation of Fe_2 is usually bacterially catalysed. In anaerobic and vertical flow wetlands, vegetation is planted in an organic substrate underlaid by limestone. Along with limestone dissolution, sulphate-reducing bacteria (SRB) are responsible for increasing the pH of the AMD in these systems, creating a more alkaline environment and causing metals to precipitate. In bioreactors microbial sulphate reduction is the main form of treatment for AMD. Bioreactors are able to treat very acidic and metal rich water through the use of SRB along with organic matter mixed with fine limestone. As well as SRB, other microbes play a key role in bioreactors, degrading organic matter and providing compounds utilised by SRB. However, little research has been carried out to determine specific species which help to establish the bioreactor system (Skousen *et al*., 2017). Sulphate-reducing bacteria found in AMD are discussed further below. As well as the use of SRB showing promise in the treatment of AMD, the removal of arsenic, one of the major contaminants frequently present in AMD, is also possible through the use of native AMD microorganisms. Three strains of bacteria from the genus *Thiomonas,* isolated from AMD, have been shown to remove arsenic from AMD through the oxidation of arsenite (As^{3+}) to arsenate (As^{5+}) which is then co-precipitated along with Fe^{3+} (Bruneel *et al*., 2003; Hallberg, 2010). Bioremoval of arsenic has also been demonstrated through adsorption on to the sulphide mineral jarosite ($KFe^{3+}{}_3(OH)_6(SO_4)_2$) which is generated by *Acidithiobacillus ferrooxidans* during the oxidation of iron sulphides (Natarajan, 2008; Asta *et al*., 2009).

As discussed above, the impact microbial life has on mining and the production and management of AMD is significant and a full understanding of the microbial

consortium present in AMD is therefore of high interest to the scientific community. A number of studies have been carried out to profile the microbial communities at numerous AMD sites; extensively studied sites include: Richmond mine (Iron Mountain, USA), the Tinto River (Spain), Cae Coch (UK), Mynydd Parys (UK), the Carnoulès (France), the Drei Kronen (Germany) the Ehrt (Germany) and the Los Rueldos (Spain) (Mendez-Garcia *et al*., 2015). The main environmental factors that influence the structure of microbial communities in AMD are pH, temperature, concentrations of dissolved metals/other solutes, total organic carbon and dissolved oxygen, however, although the microbial populations differ slightly depending on these factors, there are common findings from all sites. To date, the dominant bacterial phyla observed in AMD are *Proteobacteria, Nitrospirae, Actinobacteria, Firmicutes* and *Acidobacteria* while Archea observed in AMD are predominantly from the order Thermoplasmatales (Chen *et al.*, 2016). Key metabolic functions drive the structure of microbial AMD communities. Organic carbon and nitrogen are in limited supply in AMD environments; organisms capable of carbon and nitrogen fixation coupled with iron and sulphur oxidation are required in AMD environments to maintain a steady supply within the ecosystem (Chen *et al*., 2016). Organisms that carry out these functions are therefore vital for the maintenance of a steady microbial community and reduction in their numbers could result in the collapse of the entire microbial ecosystem (Tyson *et al*., 2005). Nitrogen fixation in these systems is likely carried out by small numbers of keystone species that are present in low abundances, including *Leptospirillium* and *Acidithiobacillus* species (Chen *et al*., 2015; Hua *et al*., 2015). Iron and sulphur oxidation are essential processes for chemolithoautotrophic acidophiles to obtain energy as well as key processes in the generation of AMD and are driven by bacteria from genera including *Leptospirillum*, *Ferrovum*, *Acidithiobacillus*, *Sulfobacillus* and the archaea *Sulfolobus* and *Ferroplasma (Johnson & Hallberg 2003; Johnson & Hallberg, 2008).* As discussed previously, SRB are important for the bioremediation of AMD, especially in bioreactors. Identifying acid-tolerant SRB that can naturally thrive in AMD is therefore of great interest. SRB are a physiologically unique group of microorganisms, which are chemoorganotrophic or chemolithotrophic, and generally active in anoxic waters (Giloteaux *et al*., 2013). Some SRB are not able to survive the harsh AMD conditions (Cabrera *et al.,* 2006). However a

number of studies have found SRB that can thrive in AMD. SRB from the genera *Desulfosporosinus, Syntrophobacter* and *Desulfurella* were found to be performing local, natural bioremediation of AMD in the Tinto river, by reducing dissolved sulphates to sulphides causing precipitation of iron and heavy metals (Sánchez-Andrea *et al*., 2012). A diverse range of SRB inhabiting ecological niches that encompassed a range of pH, temperature and chemical compositions were found in AMD from Carnoulès mine (Giloteaux *et al*., 2013). Taxonomic classification carried out using *dsrAB* (dissimilatory sulfite reductase) gene sequences found SRB affiliated to the Desulfobulbaceae, Desulfohalobiaceae and Desulfomicrobiaceae families in areas with differing pH and chemical compositions. SRB were found which were able to tolerate pH as low as 1.2 and very high levels of s*ulphate*, iron and arsenic, demonstrating the potential for utilising SRB that are naturally occurring in AMD for bioremediation purposes (Giloteaux *et al*., 2013). Colonies of SRB, identified by their deposition and accumulation of metal sulphides, were found in an acidic (pH 2.5–2.75) metal-rich stream running off an abandoned sulphide mine in Spain with 16S rRNA gene analysis indicating that they may be novel species (Rowe *et al.,* 2007).

1.3 Aims of this thesis

This thesis aims to provide a greater understanding of the microbial communities present both with the potential algal biofuel Botryococcus *braunii* and within acid mine drainage at a disused mining site in Cornwall (Wheal Jane and Wheal Maid). As previously discussed, the presence of microbial communities with microalgae can have a negative or positive effect on biomass and lipid production. A full understanding of bacteria living with microalgae may lead to the development of optimum communities of bacteria which can be co-cultured with microalgae for the biofuel industry. *B. braunii* shows great promise as a biofuel due to its exceptionally high lipid content. However, like other algal biofuels its large-scale production is hindered by issues related to growth rates and costs involved with lipid harvesting. This thesis aims to contribute further knowledge about which bacteria are living in consortia with *B. braunii* which

may in turn contribute to research in creating optimal microbial communities to aid *B. braunii* growth for the biofuel industry. Further exploration of microorganisms found living in AMD may lead to the identification of more species which can be utilised for bioremediation purposes. Additionally, variation exists between the microbial populations found in different AMD sites depending on a range of physical and geo-chemical environmental factors. It is therefore not possible to discuss what an optimum microbial community for bioremediation would look like that would be universally suited to all AMD sites. Instead it is important to look at individual sites to determine which microorganisms are likely to be thriving there. The Wheal Maid and Wheal Jane AMD sites in Cornwall have not yet been the subject of large studies characterising the microbial population, and this thesis aims to contribute to knowledge in this area. To achieve these objectives 16S rRNA, whole genome and shotgun metagenomic sequencing of the microbial consortia present with *B. braunii* and of the microbial population found in AMD was carried out. Bioinformatic methods will be applied to these sequencing sets with the aim of identifying members of the populations and gain greater insight into metabolic activity of microorganisms in these environments. Additionally, an appraisal of some of the tools available for sequence assembly and taxonomic classification will be carried out.

The aims and objectives of this thesis are to:

- use whole genome and 16S rRNA gene sequencing methods to characterise the prokaryotic microbial community associated with the oleaginous alga *Botryococcus braunii*.
- identify genes of interest (e.g those involved in symbiotic relationships) from the whole genomes of prokaryotes associated with *B. braunii.*
- use 16S rRNA and metagenomic sequencing to characterise the prokaryotic microbial community found living in acid mine drainage at the Wheal Jane mine and Wheal Maid tailings lagoon (Cornwall).
- determine if novel organisms and/or organisms with bioremediation gene pathways are present at Wheal Jane/Wheal Maid.
- comparatively evaluate bioinformatics software used in the analysis of microbial communities.

Chapter two: Genome sequencing of five bacterial strains isolated from *Botryococcus braunii* strain Guadeloupe

2.1 Introduction

2.1.1 The importance of *Botryococcus braunii* to the biofuel industry

Microalgae are a promising source of biofuels and can be used to produce a number of fuels including biodiesel, bioethanol and biomethane (Chapter 1; Singh & Gu, 2010). *Botryococcus braunii* has been identified as a species that could potentially be farmed for large scale biofuel production (Hillen *et al*., 1980; Banerjee *et al*., 2002). It is a unicellular, colonial, photosynthetic microalga belonging to the phylum Chlorophyta, and is found globally in brackish, fresh and saline water (Aaronson *et al*., 1983; Banerjee *et al*., 2002). *B. braunii* can produce and accumulate high levels of hydrocarbons (up to 70 % of its dry weight, depending on the strain), which can be converted into liquid fuels (Wolf *et al*., 1985; Yamaguchi *et al.*, 1987: Ashokkumar & Rengasamy, 2012). Furthermore, the fact that *B. braunii* can be grown in saline, fresh or waste water means that, unlike many other biofuels, there is no need to divert land from agricultural use for its production (Olguin, 2012).

Three races of *B. braunii* have been identified: A, B and L. Strains of *B. braunii* are classified as one of these three races based on the hydrocarbons that they produce. Members of A-race produce odd-numbered (C_{25}-C_{31}) alkadienes and trienes, members of B-race produce polymethylated triterpenes (C_{30}-C_{37}), whilst members of L-race produce a single tetraterpenoid hydrocarbon, lycopadiene (Metzger & Largeau, 2005) (Figure 2.1). There is a large amount of variability in the hydrocarbon content of different strains of *B. braunii,* A-race strains produce from 0.4% to 70%, B-race strains produce from 9% to 40% and L-race produce from 0.1% to 8% (of dry weight) (Metzger & Largeau, 2005). Hydrocarbons accumulate in two distinct sites within *B. braunii*: internally in cytoplasmic globules and externally in outer cell walls; this gives the advantage of making the extraction of hydrocarbons much easier than if they were stored entirely internally (Wolf *et al*., 1980). *B. braunii* forms colonies in the range of 30 μm to 2 mm, with cells immersed in an extracellular matrix of polymerised and liquid

hydrocarbons contained within an outer cell wall (Metzger *et al.* 1988). The strain of *Botryococcus braunii* used in this study (strain Guadeloupe) belongs to B-race; this strain was obtained in 2004 by the University of Exeter from Pierre Metzger at Ecole Nationale Supérieure de Chimie de Paris, France. Outside of the University of Exeter there has been little research carried out into this strain.

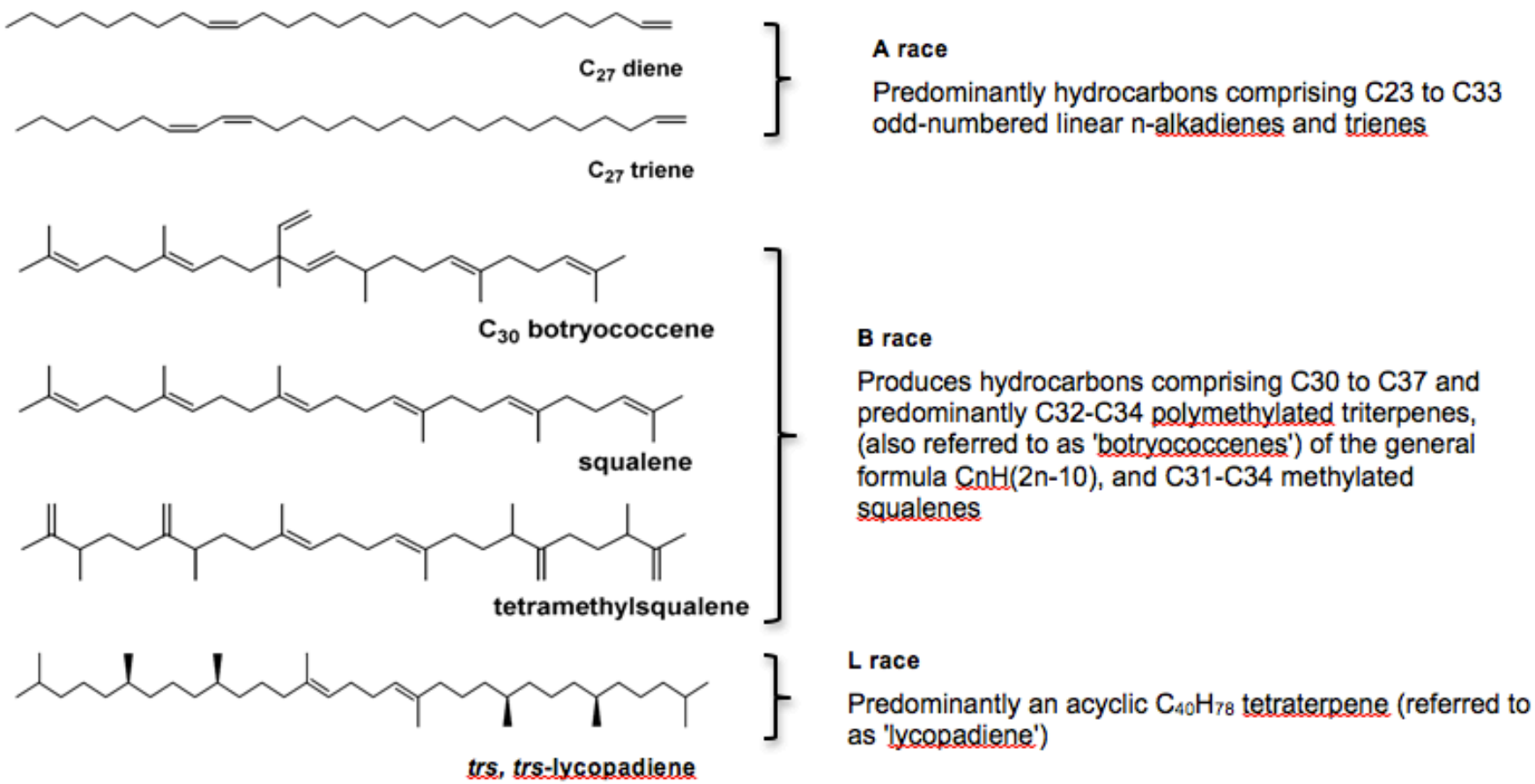


Figure 2.1 Examples of hydrocarbons produced by *Botryococcus braunii* races. (Adapted from Metzger & Largeau, 2005).

Despite the promising characteristics of *B. braunii*, there is a problem to overcome before it can be considered as a viable biofuel source and that is its slow growth rates (Tanoi *et al*., 2011). Early observations of *B. braunii* growing wild in Oak Mere, UK, demonstrated a doubling time of 24 days whilst mean generation time when growing *B. braunii* in cultures with continuous illumination was observed to be one week (Swale,1968; Belcher, 1968). However, it has since been demonstrated that a manipulation of culture conditions can reduce this generation time down to around two days with *B. braunii* growing most successfully at temperatures between 20 ^{o}C and 25 ^{o}C, with a photoperiod of twelve hours light, twelve hours dark (Qin, 2005).

2.1.2 Studies of the *B. braunii* bacterial consortium

An important factor to consider when determining the optimal growth conditions of *B. braunii* is that many strains are not axenic; instead they typically consist of a single algal strain accompanied by a variety of microorganisms living alongside *B. braunii* as both a biofilm and as planktonic populations living in the water column (Rivas *et al*., 2010). As previously discussed (Chapter 1) there are numerous ways in which bacteria can have a beneficial or detrimental effect on algal growth.

Previous studies of the *B. braunii* bacterial consortium are limited. Chirac *et al.* (1985) isolated seven bacteria from three strains of *B. braunii* (Cb, Gb and Tb) and identified them, through microscopic examination as *Arthrobacter, Corynebacterium, Pseudomonas, Erwinia, Alcaligenes, Flavobacterium* (from strain Cb) and *Streptomyces* (from strains Gb and Tb). Additionally, Chirac *et al*. took an axenic strain of *B. braunii* A and combined it in individual cultures with *Pseudomonas oleovorans*, *Corynebacterium aquatile*, *Flavobacterium aquatile*, *Azotobacter chrococcum* and *Pityrosporum ovale*. Algal biomass and hydrocarbon production were then observed for *B. braunii* A under limited and unlimited CO_2 conditions, under which *P. oleovorans* and *P. ovale* had a detrimental effect on both algal biomass and hydrocarbon yield when compared to the axenic control, *F. aquatile* caused a higher biomass and hydrocarbon yield, *A. chrococcum* increased biomass but had little effect on hydrocarbon yield and *C. aquatile* caused a slight decrease in biomass but had little effect on hydrocarbon yield. However, under limited CO_2 conditions all of the bacterial

strains appeared to have a positive effect on biomass and hydrocarbon yield, compared to the axenic control; from this the authors concluded that under such conditions any negative effects are outweighed by the benefits of CO_2 produced by bacteria. It is worth noting that within the study by Chirac *et al*. three axenic strains are discussed. Two of these 'axenic' strains (T and G) were obtained by chemical treatment or serial dilution of the non-axenic strains Gb and Tb. However, the mechanism by which axenic strain A, which was combined with the previously discussed bacterial cultures, was obtained is not clear. The authors state it came from a culture collection, but how it has been treated to be classed as axenic is not discussed. This study should therefore be treated with a little caution in relation to the claim that these individual cultures had an effect on the 'axenic' strain, as there may be other bacteria also interacting with the alga.

In order to investigate possible interactions between *B. braunii* and bacteria, Rivas *et al.* (2010) took samples of bacteria from the biofilm community of *B. braunii* LB572 (A race) and grew them on agar plates. Using 16S rRNA gene sequencing they identified three species of *Pseudomonas,* three species of *Acinetobacter,* one species of *Planomicrobium* and one species of *Rhizobium*. Following on from this, Rivas *et al.* carried out a time course study during which the microalgae cultures were inoculated with individual cultures of the bacterial species identified in the biofilm and incubated for 24 days at 20 or 25 ± 2°C with 24:0 h light dark cycle. Algal growth rates were measured; an unnamed *Rhizobium* species was the only bacterium found to significantly enhance the growth of *B. braunii*, while a strain of *Acinetobacter* sp. was observed to have a detrimental effect on the growth rate of *B. braunii*, resulting in a lower density of the alga than the control. Rivas *et al*. pointed to the fact that previous studies looking at interaction between microalgae and bacteria suggested bacteria increase algal growth due to bacterial members of the population providing essential nutrients or other unknown benefits, however the exact reason why *Rhizobium* enhanced growth in this study was not investigated.

In a study that aimed to identify *B. braunii* strains with high growth rates and high hydrocarbon production, Tanabe *et al.* (2012) isolated *B. braunii* Ba10 (B race) from a pond in South East Asia and claimed that this strain had a higher

growth rate and produced higher levels of hydrocarbons compared to those of their benchmark strain. The benchmark strain used by Tanabe *et al*. was *B. braunii* BOT-22 (B race), a strain which is known to be faster growing than many other *B. braunii* strains and which is also one of the few axenic strains of *B. braunii.* 18S rRNA analysis showed a very high level of similarity between Ba10 and BOT-22, suggesting they are very closely related, however microscopic observations showed Ba10 had numerous rod shaped bacteria present which appeared to be forming a biofilm around the rim of the Ba10 colonies. Attempts to obtain axenic cultures of these bacteria were unsuccessful using standard media for heterotrophic bacteria (trypticase soy agar, TSA) and it was suggested that this, along with their close physical association with *B. braunii* indicated a long standing mutualistic relationship rather than contamination. Furthermore, Tanabe *et al*. speculated that the presence of this bacterial symbiont was playing a role in the larger colony sizes and higher productivity of Ba10 over the axenic BOT-22, although they acknowledged the need for further investigation to test this hypothesis.

2.1.3 Aims

Whole-genome sequencing of bacteria present in the *B. braunii* consortium has not been carried out in any previous studies. Whole-genome analysis allows for genes of interest to be looked for as well as more accurate taxonomic classification to be achieved than when using 16S rRNA sequencing alone. The aim of this chapter is to determine the taxonomic and phylogenetic identity of bacteria isolated from *Botryococcus braunii*, through whole-genome sequencing, phylogenetic analysis and whole-genome comparisons. Additionally, this chapter aims to annotate and analyse these genomes in order to determine whether these bacteria contain metabolic pathways that are linked to symbiotic or pathogenic interactions with their algal host. As well as possibly enabling a better understanding of the optimal growth conditions of *B. braunii*, a better understanding of which bacteria are living alongside the microalga will contribute to current scientific knowledge of the interactions between bacterial populations and their hosts.

2.2 Materials and methods

2.2.1 Microorganisms

Botryococcus braunii race B, strain *Guadeloupe* continuously cultured since 1986 was obtained in 2004 by the University of Exeter from Pierre Metzger at Ecole Nationale Supérieure de Chimie de Paris, France. Karen Moore and colleagues in 2012 isolated cultivatable bacteria from the *B. braunii* consortium as follows: A single sample of 100 µl from a 14 day-old *B. braunii* culture was inoculated into fresh medium, diluted with 900 µl MCV and ten-fold serial dilutions performed to 10^{-7} of the original samples.,100 µl of each dilution plated on to MCV-1 % agar and LB-1 % agar plates and incubated at 25 °C for 2-7 days until bacterial colonies had formed. No evidence of *B. braunii* growth was observed on these plates. Colonies were repeatedly streaked and cultured on the appropriate medium to isolate single strains. The appearance of the bacteria was observed through microscopic examination. Table 2.1 shows the five isolates used in this study along with colony appearance determined by Moore *et al*. A single colony from each strain was cultured in 5 ml LB broth (10 g l^{-1} Bacto-tryptone, 5 g l^{-1} yeast extract, 10 g l^{-1} NaCl, pH 7.5) at 25 °C and DNA isolated from the bacteria using a Bacterial Genomic DNA Isolation Kit (Sigma, UK), according to the manufacturer's instructions.

Table 2.1 *Botryococcus braunii* consortium bacterial strains used in this study

Bacterial strain	**Colony appearance**
GCS2	Cream, uniform, smooth
GCS4	Smooth, yellow
GWS1	Smooth, white, small
SUL3	Smooth, large, cream
SUS2	Smooth, white, small

2.2.2 Library preparation and sequencing

DNA from the five bacterial strains was fragmented using sonication set at 30 s on, 90 s off, for ten min. DNA was concentrated and purified using a QIAquick column according to the manufacturer's instructions. Six genomic libraries were prepared according to the protocol in NEBNext DNA Library Prep Master Mix Set for Illumina (New England Biolabs). The libraries were amplified for a total of 8 cycles PCR, diluted 20-fold with nuclease free water, quantified and insert sizes determined, using a Bioanalyser 7500 DNA chip. 150 bp paired end sequencing, with a custom barcode, was carried out on an Illumina MiSeq. Raw paired end sequence data were uploaded to the server as FASTQ files.

2.2.3 Bioinformatic tools and software

Bioinformatic tools and websites (including references) used in this study are shown in Table 2.2.

2.2.4 Sequence assembly and quality control

Low-quality reads and adapters were removed from sequences using the fastq-mcf program from the ea-utils package.

Two *de novo* assembly algorithms were used and evaluated (see section 2.3): Velvet and SPAdes. For Velvet assemblies, paired end sequences were interleaved and VelvetOptimiser was used with a lower *k*-mer length of 81 and an upper *k*-mer length of 137.

SPAdes was run using default parameter values and the –careful option. SSPACE was used for scaffolding *de novo* assemblies, and gapfiller was used with default parameters. Reads (in FastQ format) were mapped back to the *de novo* assemblies using BWA mem. SAMtools view was used to convert SAM files into BAM files and SAMtools flagstats was used to obtain alignment statistics. Contigs of less than 200 bp were removed from the *de novo* assemblies and Quast was used to obtain assembly statistics. Following evaluation of the two assembly methods all subsequent analysis was carried out on assemblies created using SPAdes.

Benchmarking Universal Single Copy Orthologs (BUSCO) was used to assess the completeness of the genome sequence by looking for a set of genes that

are conserved across all bacteria. BUSCO was run using default parameters and the bacteria data set as the lineage. Short summaries were plotted using the BUSCO plot function.

2.2.5 Phylogenetic analysis

RNAmmer was used to extract 16S ribosomal DNA from the genome sequences assembled with SPAdes. Extracted 16S rRNA gene sequences were used as the query in a BLASTn search against the NCBI bacteria and archaea 16S ribosomal RNA sequence database, using an E-value threshold of 1×10^{-6} .Sequences with an identity of >95 % were downloaded in FastA format

Table 2.2 Software and websites used in this study

Name	Version	Available from:	Reference
bedtools	2.17.0	https://github.com/arq5x/bedtools2/releases	Quinlan & Hall,2010
BLAST executables	2.2.26	http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download	Camacho *et al*., 2009
BRIG	0.80	http://sourceforge.net/projects/brig/	Alikhan *et al.*, 2011
BUSCO	2.0	http://busco.ezlab.org/	Simão *et al*., 2015
BWA	0.7	http://bio-bwa.sourceforge.net/	Li & Durbin, 2009
Dendroscope	3.2.10	http://ab.inf.uni-tuebingen.de/software/dendroscope/	Huson *et al*., 2007
Ea-utils	1.1.2	https://code.google.com/p/ea-utils/downloads/list	Aronesty, 2013
GapFiller	1.10	http://www.baseclear.com/genomics/bioinformatics/basetools/gapfiller	Boetzer *et al*., 2012
KEGG	73.1	http://www.genome.jp/kegg/	Kanehisa *et al.*, 2000
Mauve	2.3.1	http://asap.genetics.wisc.edu/software/mauve/	Darling *et al.*, 2007
MUMmer	3.0	http://mummer.sourceforge.net/	Delcher *et al.*, 1999
MUSCLE		http://www.ebi.ac.uk/Tools/msa/muscle	Edgar, 2004
NCBI databases		http://www.ncbi.nlm.nih.gov/gquery/	
PlasmidFinder	1.3	https://cge.cbs.dtu.dk/services/PlasmidFinder/	Carattoli *et al.*, 2014
Qualimap	2.0.2	http://qualimap.bioinfo.cipf.es/	Garcia-Alcalde *et al.*, 2012
Quast	2.3	http://bioinf.spbau.ru/quast	Gurevich *et al.*, 2013
RAST	2.0	http://rast.nmpdr.org/rast.cgi	Aziz *et al*., 2008
RNAmmer	1.2	http://www.cbs.dtu.dk/cgi-bin/sw_request?rnammer	Lagesen *et al*., 2007
Spine/AGEnt/ClustAGE	0.2.1	http://vfsmspineagent.fsm.northwestern.edu/index_age.html	Ozer *et al*., 2014
SAMtools	0.1.19	http://samtools.sourceforge.net	Li *et al.,* 2009
SeaView	4.5.3	http://doua.prabi.fr/software/seaview	Gouy *et al*., 2010
SPAdes	3.5.0	http://bioinf.spbau.ru/spades	Bankevich *et al.*, 2012
SSPACE	2.0	http://www.baseclear.com/genomics/bioinformatics/basetools/SSPACE	Boetzer *et al*., 2011
Velvet	1.2.10	https://www.ebi.ac.uk/~zerbino/velvet/	Zerbino & Birney, 2008
Velvet Optimiser	2.2.5	http://bioinformatics.net.au/software.velvetoptimiser.shtml	Gladman & Seemann, 2009

and aligned using MUSCLE, with poorly aligned and divergent regions eliminated using Gblocks. Maximum likelihood phylogenetic trees were constructed using PhyML within the SeaView package, using the GTR substitution model and bootstrapping of 100. All trees were edited using Adobe Acrobat Pro and Dendroscope.

Genes for multilocus sequence analysis (MLSA) were extracted from the genome sequences by using FASTA files of gene sequences (obtained from the NCBI database) as the query sequences and draft genome sequences as the databases in a BLAST alignment, using BLASTn with default parameter values and an E-value threshold of 1×10^{-6} . Housekeeping gene sequence alignments and phylogenetic trees were constructed using the methods detailed above for 16S rRNA phylogenetics.

2.2.6 Annotation

Assembled genomes were uploaded to RAST (Rapid Annotation using Subsystem Technology) with parameters set to: domain bacteria, genetic code 11, classic RAST annotation scheme, RAST gene caller, FIGfam version 70, automatically fixed errors, fixed frame shifts, backfilled gaps, no reserved gene calls.

RAST performs annotations and assigns functional roles to proteins through the use of GLIMMER and the SEED protein database. Functional roles are then assigned to groups where together they implement specific biological processes or structural complexes. RAST refers to these groups as “Subsystems” (Aziz *et al*, 2008).

2.2.7 Sequence comparisons

Genome sequences of organisms identified as being phylogenetically close to the bacterial strains were downloaded from the NCBI Genome database for whole-genome alignments. Contigs within draft genome sequences were re-ordered using Mauve against a reference genome. BRIG (BLAST Ring Image Generator) was used to generate genome alignment images, using an E-value threshold of 1×10^{-6} . BWA mem was used to align bacterial strains GCS2, GCS4, GWS1 SUL3 and SUS2 against reference genomes. SAM to BAM file

conversions were carried out using SAMtools view. BWA alignments and GFF files (generated by RAST) were analysed using BedTools Coverage to identify differences in gene content, using default parameters.

Nucmer (part of the MUMmer package) was used to align sequences and generate average nucleotide identities using default parameters and dnadiff was used to generate report files.

Spine and AGEnt were used to compute core and accessory genomes using default parameters.

All BLAST searches were carried out using BLASTn, default parameters and an E-value threhsold of $1x10^{-6}$ Blast databases were formatted using BLASTdb and default parameters.

2.3 Sequencing and genome assembly of five bacterial strains, isolated from *Botryococcus braunii*

2.3.1 Comparison of genome assemblies generated by two different methods: Velvet and SPAdes

Whole genome sequencing of bacteria present in a *B. braunii* consortium has not been carried out in any previous studies and allows for numerous genes of interest to be looked for as well as more accurate taxonomic classification to be achieved than when using 16S rRNA sequencing alone. In order to achieve this, genomic DNA was extracted from five bacterial strains isolated from *B. braunii* race B strain Guadeloupe. Whole genome sequencing of these five bacterial isolates was carried out on an Illumina MiSeq. The MiSeq sequencer is recommended as an appropriate sequencing platform for small *de novo* genomes, with a lower cost, longer reads and faster throughput than other systems, such as the HiSeq (Illumina, 2015).

The accurate assembly of next-generation sequencing data in order to correctly reconstruct genomes is a rapidly advancing area, with new assembly tools and algorithms being constantly developed (Utturkar *et al.*, 2014). Currently, there are two widely used classes of algorithms used in assembly tools: overlap layout consensus (OLC) and de Bruijn graph. Sanger sequencing assemblers used the OLC method, however the advent of next generation sequencing brought about the development of assemblers which utilised De Bruijn graphs, which could handle the typically short reads produced by platforms such as Illumina (Li *et al*., 2012; Compeau *et al*., 2011). Figure 2.2 demonstrates the principles behind de Bruijn graph assembly. In this study two widely used assembly algorithms, Velvet and SPAdes, were used in order to compare different methods and evaluate which is best suited to this data set. These two assemblers were chosen for the following reasons: Velvet is a well-established assembler which has been used in a large number of studies for for the *de novo*

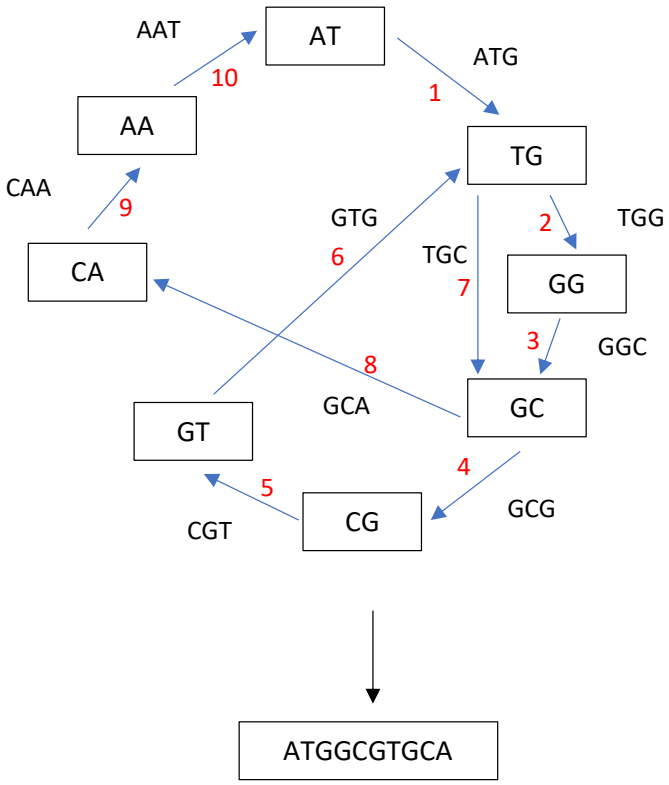


Figure 2.2 A simple example of assembly, using a de Bruijn graph, of sequence ATGGCGTGCA. The sequence has been split into *k*-mers of *k* =3. The de Bruijn graph is composed of nodes and edges; all *k*-mer prefixes and suffixes are represented as nodes (shown here as boxes) and edges represent overlapping *k*-mers. (Compeau *et al*., 2011)

assembly of whole genomes (Narzisi *et al*., 2011); however, the more recently developed SPAdes has been shown to be a promising new tool which, according to its developers, has performed better than Velvet and a range of other commonly used *de novo* assemblers, including SOAPdenovo, Abyss, Cabog, Mira and SGA (Magoc *et al.*, 2013). The Velvet set of assembly algorithms was developed in 2008 for the *de novo* assembly of very short read (25-50 bp) data sets (Zerbino and Birney, 2008). The SPAdes assembler was developed in 2012 with the dual purpose of being able to assemble both single-cell and standard (multicell) data sets (Bankevich *et al.*, 2012). The two assemblers have a number of similarities: both Velvet and SPAdes break reads into k-mers, which are then used to build *de Bruijn* graphs. Errors are removed from these graphs and optimal paths are found through them. However, whereas Velvet requires the optimisation of a number of parameters, including k-mer length, prior to genome assembly, SPAdes builds k-mer optimisation into its assembly pipeline, performing assemblies with small, sensitive k-mer sizes, as well as large, specific, k-mer sizes before merging all results into an optimal final assembly. Additionally, the SPAdes pipeline carries out error checking both before and after assembly and studies have shown final genome assemblies created using SPAdes generally have fewer errors and a higher N50 than those produced using Velvet (Magoc *et al.*, 2013; Utturkar *et al.*, 2014).

In order to compare the performance of Velvet and SPAdes assembly algorithms, Illumina genome sequence data for all five bacterial isolates were assembled with each one and key assembly statistics were compared. Tables 2.3 and 2.4 show assembly statistics for each of the five bacterial strains after initial assembly with SPAdes and Velvet and after further scaffolding. Although Velvet was consistent in producing assemblies with fewer scaffolds, the N50 was higher using SPAdes for all assemblies other than for bacterial strain SUL3, where the Velvet assembly had a marginally higher N50 of 358844 compared to 358737 (Table 2.4) (the N50 of an assembly means that 50% of bases within the assembly are contained in contigs or scaffolds equal to or larger than this value). Additionally, fewer Ns were introduced in the assemblies using SPAdes compared to Velvet. Figure 2.3, plotted using Quast, shows the cumulative lengths of assemblies, and it can be seen that, for all but SUL3, SPAdes produces larger contigs, which make up the main bulk of the assembly.

Table 2.3 Comparisons between Velvet and SPAdes genome assemblies (after removal of contigs <200 bp)

	Velvet				SPAdes			
Bacterial strain	**No. of scaffolds**	**N50**	**Sum of BP**	**Number of N's**	**No. of scaffolds**	**N50**	**Sum of BP**	**Number of N's**
GCS2	40	709958	6215910	31261	82	598826	6201495	616
GCS4	13	486744	3662927	2795	140	829987	3693076	0
GWS1	243	104519	7030114	31884	346	160707	7052010	131
SUL3	123	359317	6047182	20258	184	263234	6043086	44
SUS2	208	183348	7024799	16073	235	264748	7041690	86

Table 2.4 Comparisons between Velvet and SPAdes assemblies following additional scaffolding using Sspace and gap-filling using Base clear gap filler

	Velvet				SPAdes			
Bacterial strain	**No. of scaffolds**	**N50**	**Sum of BP**	**Number of N's**	**No. of scaffolds**	**N50**	**Sum of BP**	**Number of N's**
GCS2	28	709720	6212176	3231	74	854972	6201055	52
GCS4	12	486606	3661630	14	139	944354	3693077	11
GWS1	194	135302	7015044	1746	316	237382	7051455	361
SUL3	90	358844	6037213	2189	146	358737	6041380	167
SUS2	166	183921	7012715	935	208	308460	7040134	73

Table 2.5 Percentage of raw sequencing reads mapped back to *de novo* assemblies constructed using Velvet and SPAdes

Strain	**Velvet (%)**	**SPAdes (%)**
GCS2	99.64	99.65
GCS4	99.76	99.80
GWS1	99.63	99.72
SUL3	99.71	99.73
SUS2	99.75	99.77

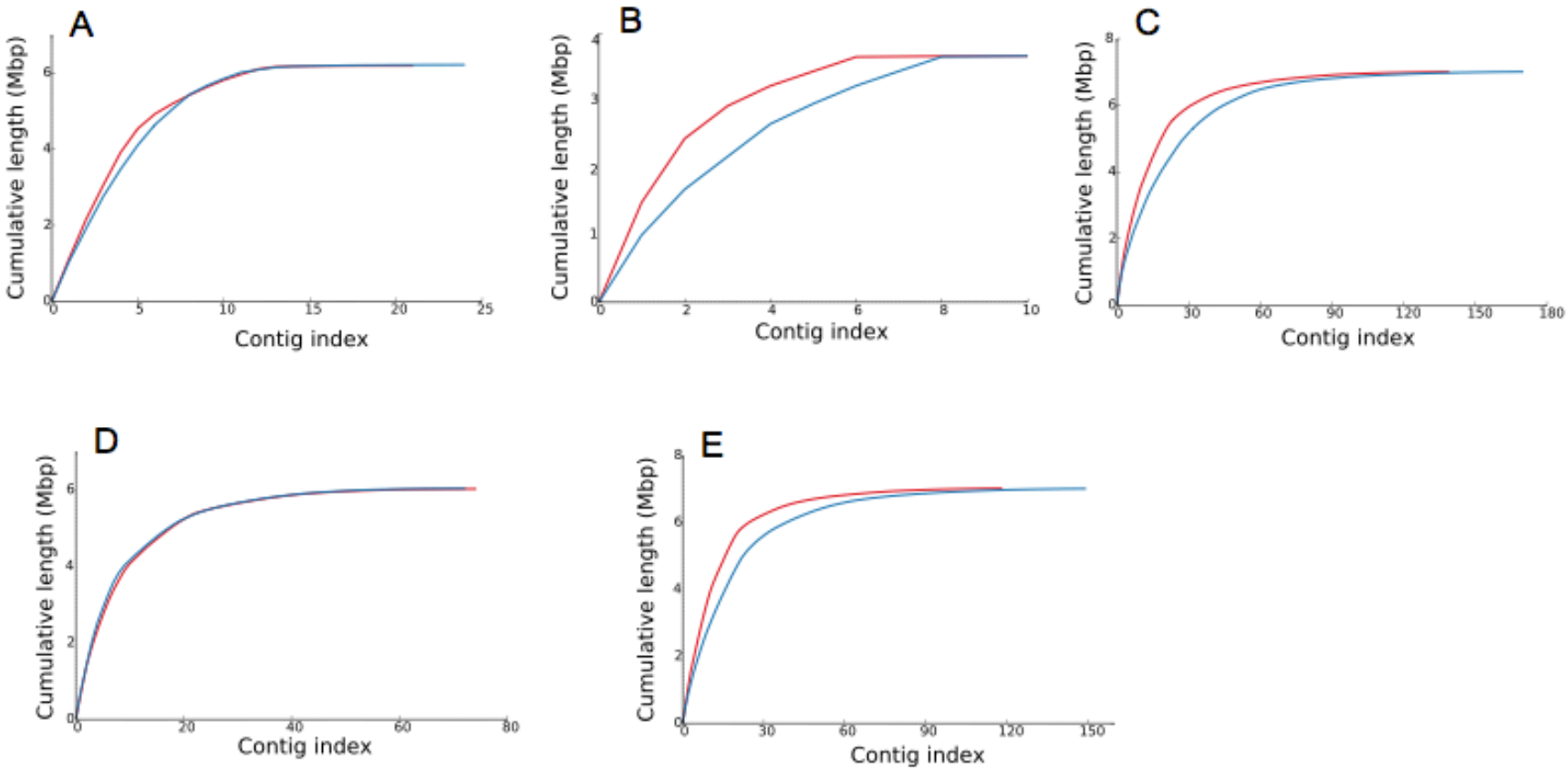


Figure 2.3 Cumulative length plots. On the x-axis, contigs are ordered from largest to smallest. The y-axis shows the cumulative length of the assembly. Contigs >500 bp are included. Red indicates SPAdes assemblies and blue indicates Velvet assemblies for: A = GCS2, B = GCS4, C = GWS1, D = SUL3, E = SUS2.

In order to determine the number of reads being included in the *de novo* assemblies, the raw Illumina reads were aligned against the *de novo* assemblies using the Burrows Wheeler Aligner (BWA mem). The results from the BWA alignment (Table 2.5) show a high proportion (over 99%) of the Illumina reads have been included in the *de novo* assemblies, with marginally more being included in the SPAdes assemblies than the Velvet assemblies. Following on from this evaluation of the two assembly methods it was decided to use the SPAdes assemblies for all further analysis due to their generally higher N50 and lower N content. All subsequent analysis in this chapter has been carried out on these SPAdes assemblies.

2.3.2 Assessing the completeness of the genome using BUSCO

Having assembled the five genomes using SPAdes, the tool Benchmarking Universal Single Copy Orthologs (BUSCO) was then used in order to assess the completeness of the genome based on gene content. BUSCO looks for sequences encoding 148 BUSCOs that are conserved across all bacterial species and which should only occur as a single copy within the genome. Additionally, BUSCO reports on whether any BUSCOs are missing, duplicated or fragmented. Figure 2.4 shows output from BUSCO. The results from BUSCO were identical when run on both the SPAdes and Velvet assemblies.

All of the draft genomes had fewer than 5 % of BUSCOs missing, with 4 out of 148 missing BUSCOs (2.7 %) in GCS2, GWS1 and SUS2, 6 missing from SUL3 (4 %) and 7 missing BUSCOs (4.7 %) from GCS4. One BUSCO was fragmented, in GCS2 and GWS1 and a maximum of 2 BUSCOs were duplicated per genome. Comparisons of numbers of missing BUSCOs between the *B. braunii* bacterial strains and the genomes of bacterial strains they are closely related to are documented in subsequent sections, where it can be seen that these numbers appear average for these bacterial genera. Therefore, it can be concluded that the genomes that have been assembled from the *B. braunii* bacteria are of a high enough quality to further analyse.

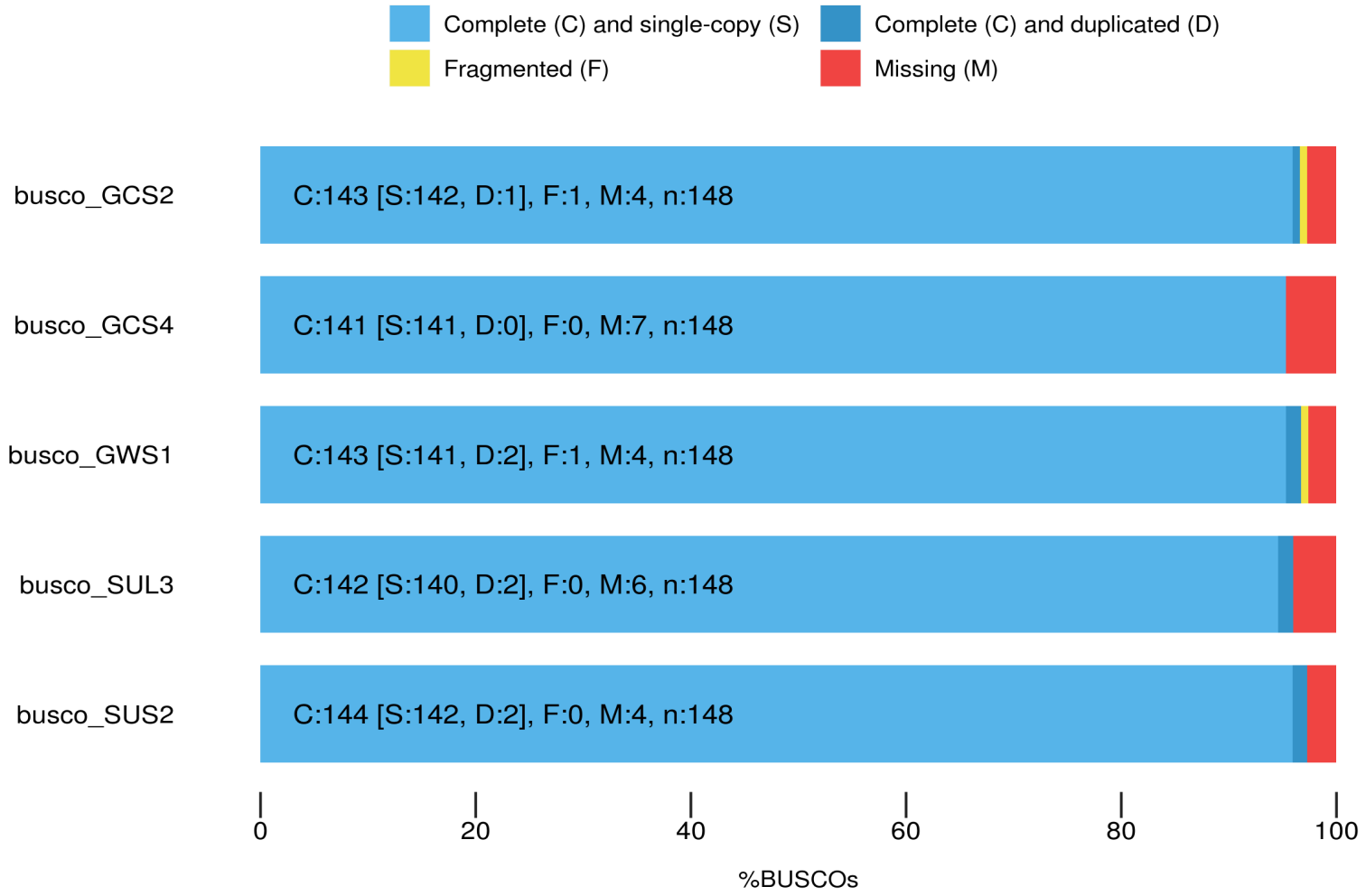


Figure 2.4 Output from BUSCO, created using the BUSCO bacterial data set to search within the five draft genomes for genes conserved across all bacterial species. The legend is shown above the plots and figures are given on each plot.

2.4 Bacterial isolate GCS2 belongs to species *Achromobacter piechaudii.*

2.4.1 Phylogenetic analysis of bacterial isolate GCS2

To identify bacterial isolate GCS2 its 16S rRNA gene sequence was extracted from the assembled genome using RNAmmer. Due to its extensive use in a wide range of studies investigating bacteria obtained from both clinical and environmental samples, as well as its popularity as a method for identifying rare or uncultivable bacteria, there is now a very large database of 16S rRNA gene sequences. Numerous bioinformatics tools have been developed in order to analyse unknown 16S rRNA gene sequences (Drancourt *et al*., 2000). In order to identify organisms similar to bacterial strain GCS2 the extracted 16S rRNA gene sequence was used in a BLAST search against the NCBI Nucleotide database. 16S rRNA gene sequences identified through BLAST as having at least 95% identity to bacterial strain GCS2 were then used to create a maximum likelihood phylogenetic tree (Figure 2.5). This clearly shows bacterial strain GCS2 falls in a clade with species from the genus *Achromobacter*.

Despite its widespread use, studies have shown phylogenetic analysis using 16S rRNA gene sequences alone often have poor resolution at the species level (Janda and Abbott, 2007). To gain better resolution, multilocus sequence analysis using housekeeping genes is often used in phylogenetic studies. It has been found that the higher degree of sequence divergence of housekeeping genes is superior for identification purposes and has better discriminatory power than 16S rRNA analysis alone (Case *et al*., 2007: Martens *et al*., 2008), although which range of housekeeping genes is best suited to phylogenetics is still not agreed upon (Carrasco, 2013). Gene sequences for *atpD, recA, rpoB* and *tyrB* were extracted from the assembled genome of bacterial strain GCS2; these genes were chosen due to their use in a previous study by Ridderberg *et al*. (2012) which carried out multilocus sequence analysis of *Achromobacter*

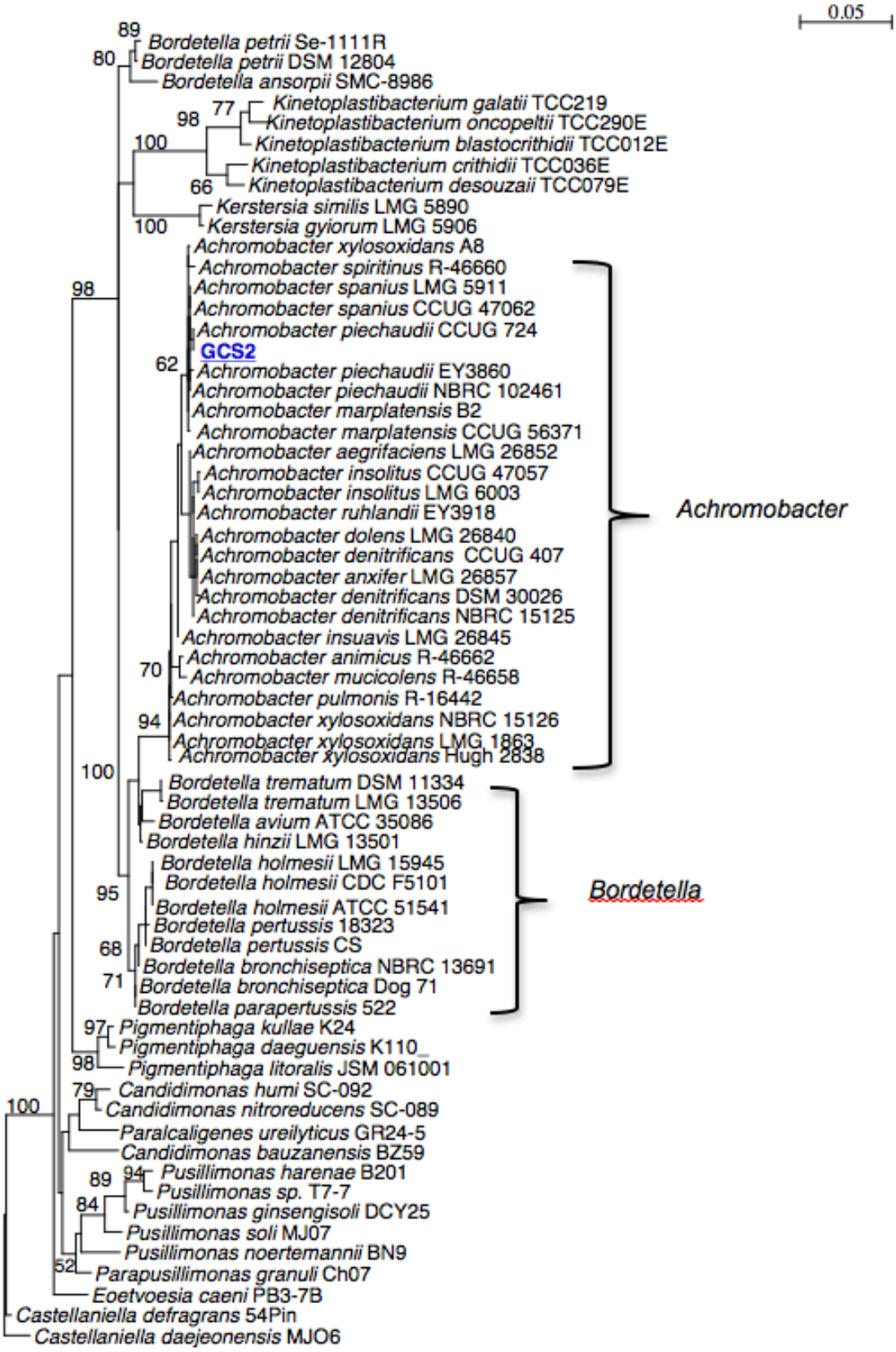


Figure 2.5 Phylogram constructed from maximum likelihood analysis (PhyML) of 16S rRNA gene sequence data for bacteria, identified through BLAST as having > 95% identity to the 16S rRNA gene sequence of bacterial strain GCS2. Bootstrapping was set at 100 and values are shown for each node with > 50% bootstrap support. The tree is rooted with *Castellaniella daejeonensis* MJO6 as the outgroup.

isolates from clinical samples. Gene sequences for *atpD, recA, rpoB* and *tyrB* were obtained for all *Achromobacter* species where available in the NCBI database. These gene sequences were then concatenated and used to create a maximum likelihood phylogenetic tree (Figure 2.6) which indicates sample GCS2 is a strain of *Achromobacter piechaudii*.

Having concluded from phylogenetic evidence that bacterial strain GCS2 is most likely to be a strain of *A. piechaudii,* the average nucleotide identity (ANI) of GCS2 and *A. piechaudii* ATCC 43553 was calculated (using MUMmer) at 98.4%. An ANI cutoff score of >95% is generally used to indicate two genomes as belonging to the same species (Figueres *et al.*, 2014). Therefore, in all subsequent sections bacterial strain GCS2 will be referred to as *A. piechaudii* GCS2.

Achromobacter piechaudii is a Gram negative, oxidase positive member of the Alcaligenaceae family (Betaproteobacteria, Burkholderiales) (Kay *et al.*, 2001). *A. piechaudii* has previously been isolated from a variety of environments, including clinical samples, soil and aquatic habitats (Ronen *et al*., 2005, Schoch and Cunha, 1988). A strain of *Achromobacter piechaudii* has been identified as an alkane degrader; *A. piechaudii* strain 01 was isolated from petroleum reservoir waste water in Iran where it was found to be contributing to biodegradation (Hassanshahian *et al*., 2013). Subsequent PCR screening identified the presence of the *alkB* gene in *A. piechaudii* strain 01. The *alkB* gene codes for an enzyme involved in alkane degradation pathways and has been subsequently looked for in *A. piechaudii* GCS2 (section 2.10), however it was not found. The only sequence data available for *A. piechaudii* strain 01 is a partial 16S rRNA gene sequence so genome comparisons between this alkane degrading strain and strain GCS2 have not been carried out. There is no evidence in the current literature to suggest *A. piechaudii* forms any kind of symbiotic or pathogenic relationship with plants or algae, although a 2009 study by Dimkpa *et al.* demonstrated an increased tolerance of salt stress in tomato plants inoculated with *A. piechaudii*. This study will analyse the genome of *A. piechaudii* GCS2 to determine any key differences between *A. piechaudii* GCS2 and other *A. piechaudii* strains as well as determining if this strain has any features which would indicate it is interacting with *B. braunii* .

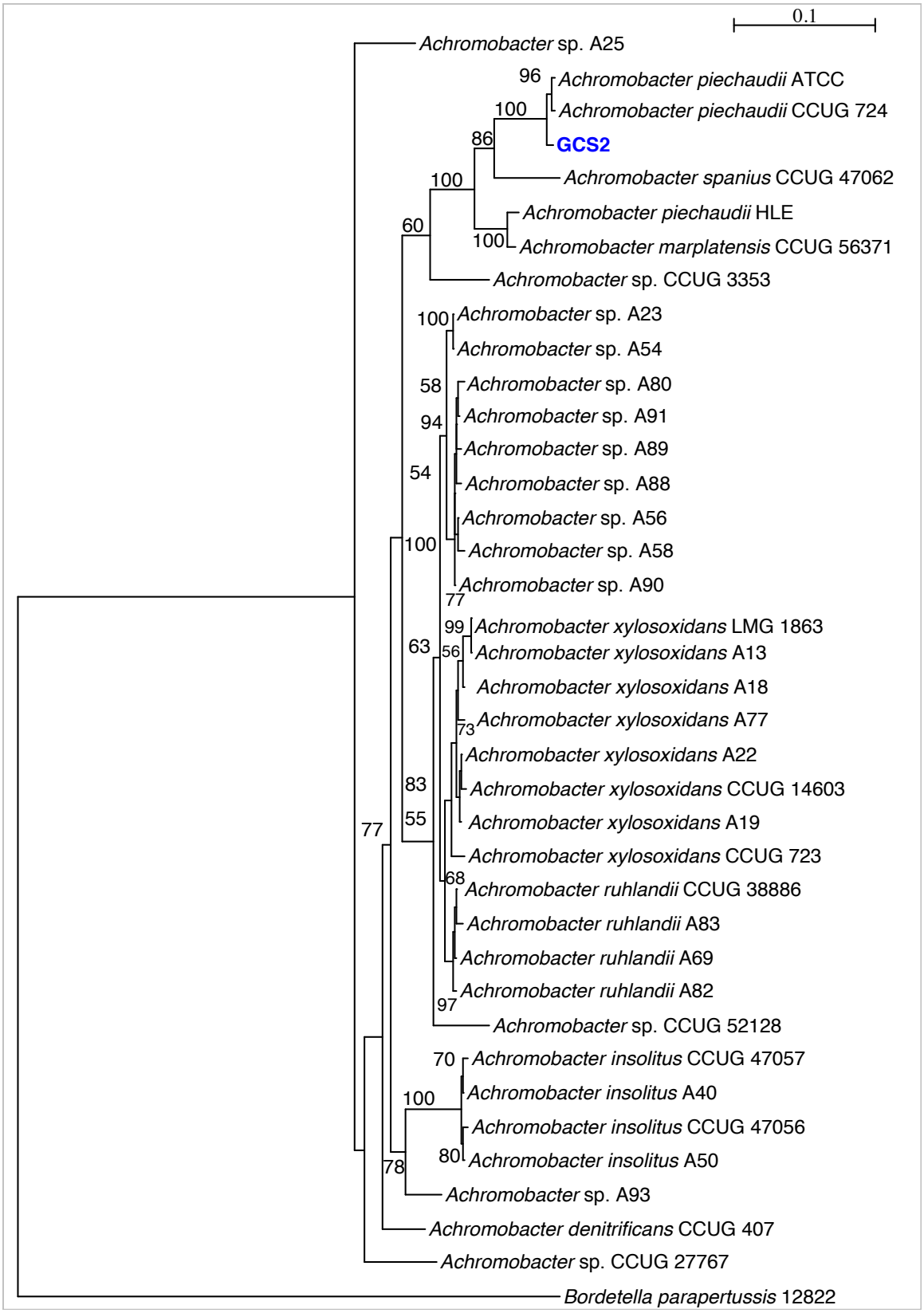


Figure 2.6 Phylogram constructed from maximum likelihood analysis (PhyML) of *atpD*, *recA*, *rpoB* and *tyrB* sequence data for bacterial strain GCS2 and *Achromobacter* species. Bootstrapping was set at 100 and values are shown for each node with > 50% bootstrap support. The tree is rooted with *Bordetella parapertussis* 12822 as the outgroup.

2.4.2 Whole genome comparisons of *A. piechaudii* GCS2 with previously sequenced strains of *Achromobacter* identify differences.

In order to determine how similar *A. piechaudii* GCS2 is to other strains of *Achromobacter* as well as to determine if there are genes unique to *A. piechaudii* GCS2 whole genome comparisons were carried out between *A. piechaudii* GCS2 and all *Achromobacter* genomes available from the NCBI database (Supp. Table 1). It should be noted that *A. piechaudii* CCUG, which appears phylogenetically close to GCS2, does not have an available genome sequence.

Whole genome comparisons between *A. piechaudii* GCS2 and other *Achromobacte*r genome sequences using the BLAST Ring Image Generator (BRIG) demonstrate that *A. piechaudii* GCS2 is most similar to *A. piechaudii* ATCC 43553 (Figure 2.7). However, there are also clear differences between the two sequences, most significantly on scaffolds 1, 2, 3, 4, 5, 8, and 9. To further investigate these differences, the raw reads from *A. piechaudii* GCS2 were aligned against the other *Achromobacte*r genome sequences using BWA mem. The depth and breadth of coverage of features in *A. piechaudii* GCS2 by the *Achromobacter* reference genomes were computed using BedTools. The greatest depth and breadth of coverage was achieved by *A. piechaudii* ATCC 43553 , which is in keeping with the results from BRIG analysis.

Achromobacter piechaudii ATCC is an isolate from a clinical sample (a nose wound) and part of the Human Microbiome Project (HMP) (http://www.hmpdacc.org/). Both *A. piechaudii* ATCC 43553 and *A. piechaudii* GCS2 have a GC content of 64 % and have genome sizes of 6.1Mb. On submission to the NCBI genome database strain ATCC has 5577 predicted genes assigned to it whilst strain GCS2 has 5544. As discussed in section 2.3.2 *A. piechaudii* GCS2 has four missing BUSCOs; *A. piechaudii* ATCC was also analysed with BUSCO and found to have five missing BUSCOs, four of which were the same as those missing from *A. piechaudii* GCS2.

The majority of genes present in *A. piechaudii* GCS2 but absent from *A. piechaudii* ATCC 43553 encode hypothetical proteins or phage related proteins.

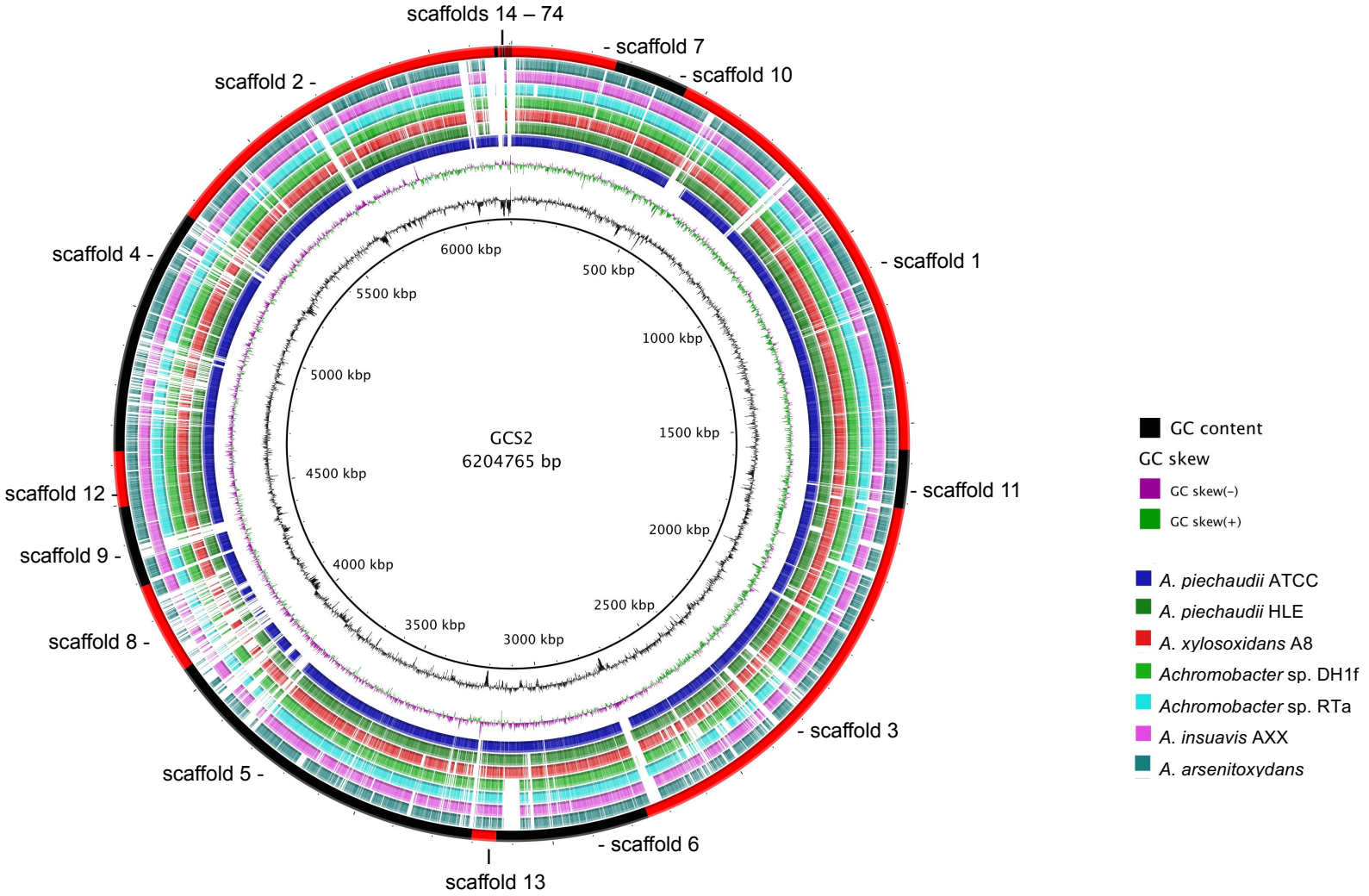


Figure 2.7 Whole-genome comparisons between *A. piechaudii* GCS2 and seven other *Achromobacter* species created using the Blast Ring Image Generator (BRIG). The alternating red and black circle indicates scaffolds of *A. piechaudii* GCS2, whilst inner coloured circles give a graphical representation of areas of homology between the reference sequence *(A. piechaudii* GCS2) and the query sequences (other *Achromobacter* species).

However, genes involved in the type VI secretion system are also present in *A. piechaudii* GCS2 but absent from *A. piechaudii* ATCC 43553. These are *icmF, impFBGACJDH, vasAEB*, and *vgrG* and are all on scaffold one. Further discussion of this type VI secretion system present in *A. piechaudii* GCS2, is detailed later (Section 2.9). Encoded on scaffold five of *A. piechaudii* GCS2 but also absent from *A. piechaudii* ATCC is a cobalt/zinc/cadmium resistance protein as well as a probable Co/Zn/Cd efflux system membrane fusion protein, indicating resistance to copper, zinc and cadmium in *A. piechaudii* GCS2.

Alignment of *A. piechaudii* GCS2 against *A. piechaudii* ATCC using BWA mem resulted in 90.0% of strain GCS2 reads being mapped. Using Qualimap, regions of *A. piechaudii* ATCC that had no coverage by *A. piechaudii* GCS2 were identified and features of interest looked for. Present on contigs within *A. piechaudii* ATCC, which had a zero level of coverage by *A. piechaudii* GCS2, were a number of hypothetical proteins and phage-related genes, as well as a twitching motility protein and type II/IV secretion system proteins. Both the type II and type IV secretion systems enable the transport of cytoplasmic proteins across the cell envelope of Gram negative bacteria and are often associated with pathogenesis (Sandkvist, 2001). Twitching motility is also important in Gram negative pathogenic bacteria, enabling the colonisation of plant or animal hosts (Mattick, 2002). The presence of these pathogenic features in *A. piechaudii* ATCC but not *A. piechaudii* GCS2 indicates they have different adaptations for the different hosts they have been found on.

Genes involved in vitamin synthesis, secretion systems and nitrogen fixation were also looked for within the genome sequence of *A. piechaudii* GCS2; these are discussed in more detail in sections 2.8 – 2.10

2.5 Bacterial isolates GWS1 and SUS2 are very closely related and fall within the genus *Shinella*

2.5.1 Bacterial strains GWS1 and SUS2 are members of the genus *Shinella*

As with bacterial strain GCS2, the first step in determining the taxonomy of bacterial strains GWS1 and SUS2 was to extract 16S rRNA gene sequences using RNAmmer from the assembled genomes. Alignment (using MUMmer) of the 16S rRNA gene sequences from GWS1 and SUS2 showed that the two sequences were 100% identical and therefore only one of these sequences was used in subsequent analysis. The 16S rRNA gene sequence from GWS1/SUS2 was used as the query sequence in a BLASTn search against the NCBI 16S ribosomal RNA sequences (bacteria and Archaea) database. A maximum likelihood phylogenetic tree was constructed from the 16S rRNA gene sequence from bacterial strains SUS2 / GWS1 along with 16S rRNA gene sequences which had >95% identity to SUS2 / GWS1 according to the BLAST search (Figure 2.8). The 16S rRNA phylogeny includes bacteria belonging to a number of families: Rhizobiaceae, Beijerinckiaceae, Brucellaceae Phyllobacteriaceae and one member of Micrococcaceae. All of these are from the order Rhizobiales (class: Alphaproteobacteria), with the exception of Micrococcaceae. Bacterial strain GWS1 / SUS2 is in a clade with members of the *Shinella* genus. Additional phylogenetic analysis was carried out using the housekeeping genes *recA* and *atpD*; these genes were chosen as they were previously used in a study by Gaunt *et al*. (2001) for phylogenetic analysis of a number of rhizobia. A maximum likelihood tree was constructed from the *recA* and *atpD* genes extracted from the genomes of bacterial strains GWS1 and SUS2 along with *recA* and *atpD* sequences from all available *Shinella* species on the NCBI database and a range of bacteria belonging to the order Rhizobiales (Figure 2.9). The *recA* and *atpD* tree placed bacterial strains GWS1 and SUS2 in a clade with members of the *Shinella* genus adding further evidence that bacterial strains GWS1 and SUS2 are strains of *Shinella* and they

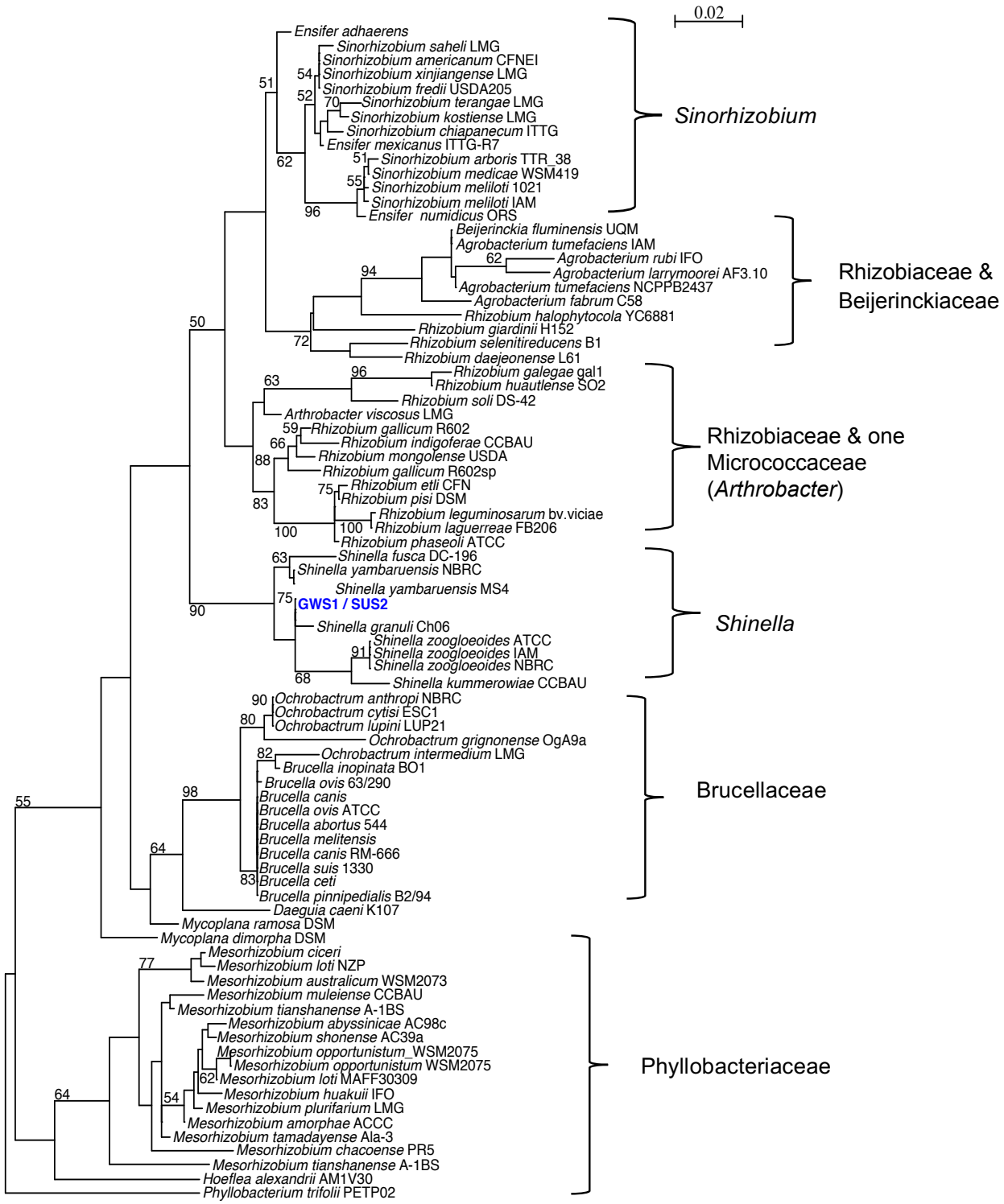


Figure 2.8 Phylogram constructed from maximum likelihood analysis (PhyML) of 16S rRNA gene sequence data for bacteria, identified through BLAST as having > 95% identity to 16S rRNA gene sequences of bacterial strains GWS1 and SUS2. Bootstrapping was set at 100 and values are shown for each node with > 50% bootstrap support. The tree is rooted with *Phyllobacterium trifolii* PETP02.

Figure 2.9 Phylogram constructed from maximum likelihood analysis (PhyML) of *atpD* and *recA*, sequence data for bacterial strains GWS1, SUS2 and bacteria from the family Rhizobiaceae. Bootstrapping was set at 100 and values are shown for each node with > 50% bootstrap support. The tree was rooted with *Bordetella parapertussis* 12822.

will therefore be referred to as *Shinella* sp. GWS1 and *Shinella* sp. SUS2 in all subsequent sections.

To further refine the phylogenetic relationship between *Shinella* spp. GWS1/SUS2 and other members of the *Shinella* genus phylogenetic trees were constructed using 16S rRNA gene sequences, *recA* and *atpD* genes as before, however only *Shinella* was included (Figures 2.10 and 2.11). Additionally, this allowed for the inclusion of *Shinella* sequences which were deposited in the NCBI database after the initial phylogenetic analysis had been carried out. Both of these trees place *Shinella* spp. SUS2/GWS1 with *Shinella* sp. DD12, a strain which is discussed below.

Shinella is a relatively new genus, having been introduced in 2006 with the reclassification of *Zoogloea ramigera* to *Shinella zoogloeoides* and the isolation of *S. granuli* and *S. zoogloeoides* from an up-flow anaerobic sludge blanket reactor and the activated sludge of a cooking wastewater treatment plant respectively (An *et al.*, 2006). In 2008 *Shinella kummerowia* was isolated from root nodules of *Kummerowia stipulacea*, a plant in the legume family, in China. *S. kummerowia* was described as symbiotic and differed from the two previously discovered strains of *Shinella*, *S. granuli* and *S. zoogoleoides*, in that it had *nodD*, *nodC* and *nifH* present, all of which are key genes in nitrogen fixation and are generally found in rhizobia (Lin *et al*, 2008). In 2009 *Shinella yambaruensis* was isolated from soil in Japan (Matsui *et al*., 2009). 16S rRNA analysis indicated *S. yambaruensis* was most closely related to *S. granuli* and *S. zoogloeoides*, furthermore *S. yambaruensis* along with the other *Shinella* species could be differentiated from their nearest phylogenetic neighbours by their utilisation of various sugars and sugar alcohols. However, unlike other *Shinella* species, *S. yambaruensis* showed no motility and could not grow at pH 10 (Matsui *et al.,* 2009). In 2010 and 2011 two more *Shinella* species were discovered: *Shinella fusca* and *Shinella daejeonensi*s. *S. fusca* was isolated from domestic waste compost (Vaz-Moreira *et al*.,2010) and *S. daejeonensi*s was isolated from sludge of a leachate treatment plant (Lee *et al*., 2011). Like *S. granuli* and *S. zoogloeoides, S. fusca* and *S. daejeonensis* both lacked the *nifH* gene and did not appear to have nitrogen fixing properties. In 2016, the first whole genome of a member of the *Shinella* genus was sequenced: *Shinella*

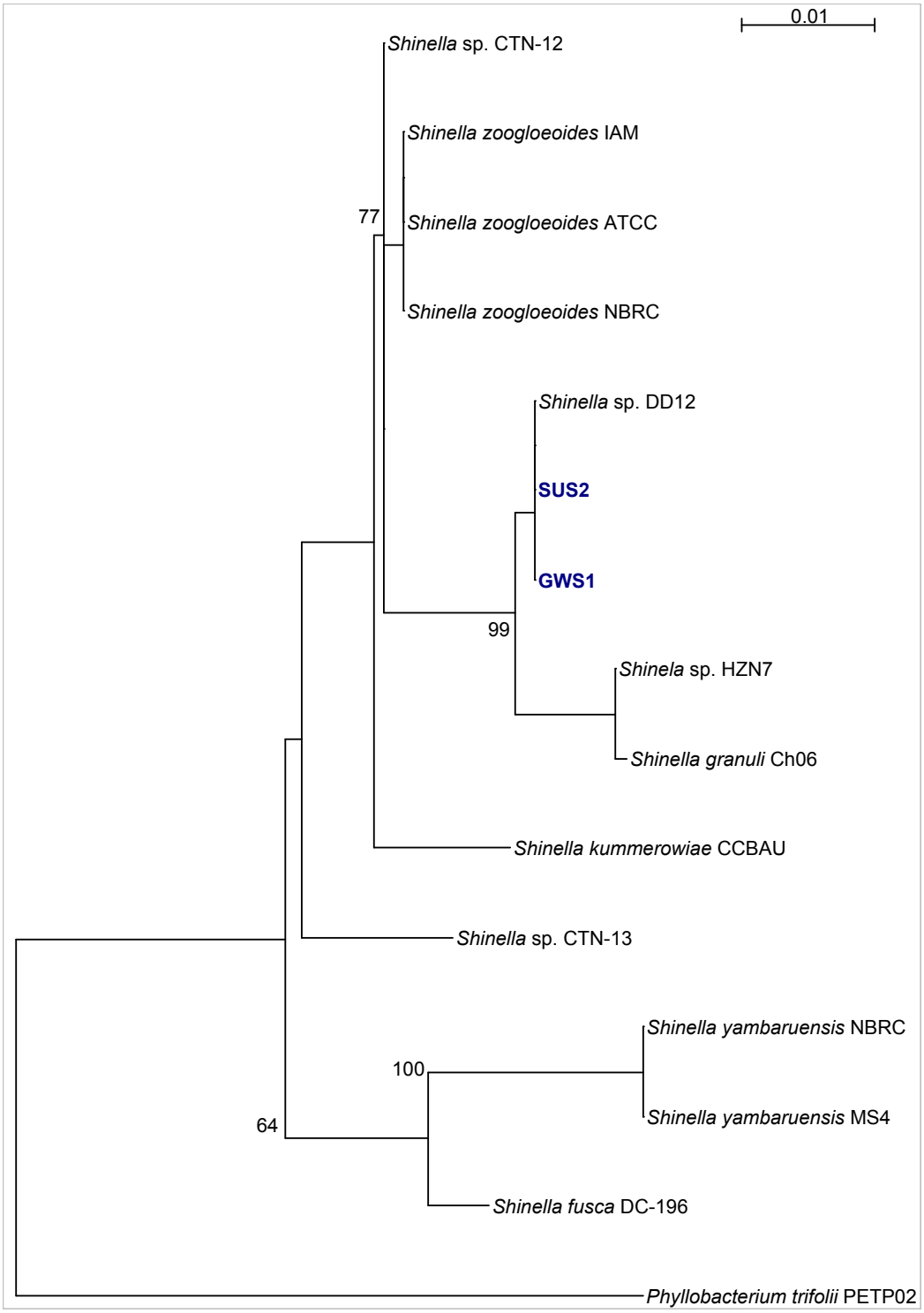


Figure 2.10 Phylogram constructed from maximum likelihood analysis (PhyML) of 16S rRNA gene sequence data for bacterial strains GWS1, SUS2 and a range of *Shinella* species. Bootstrapping was set at 100 and values are shown for each node with > 50% bootstrap support. The tree is rooted with *Phyllobacterium trifolii* PETP02.

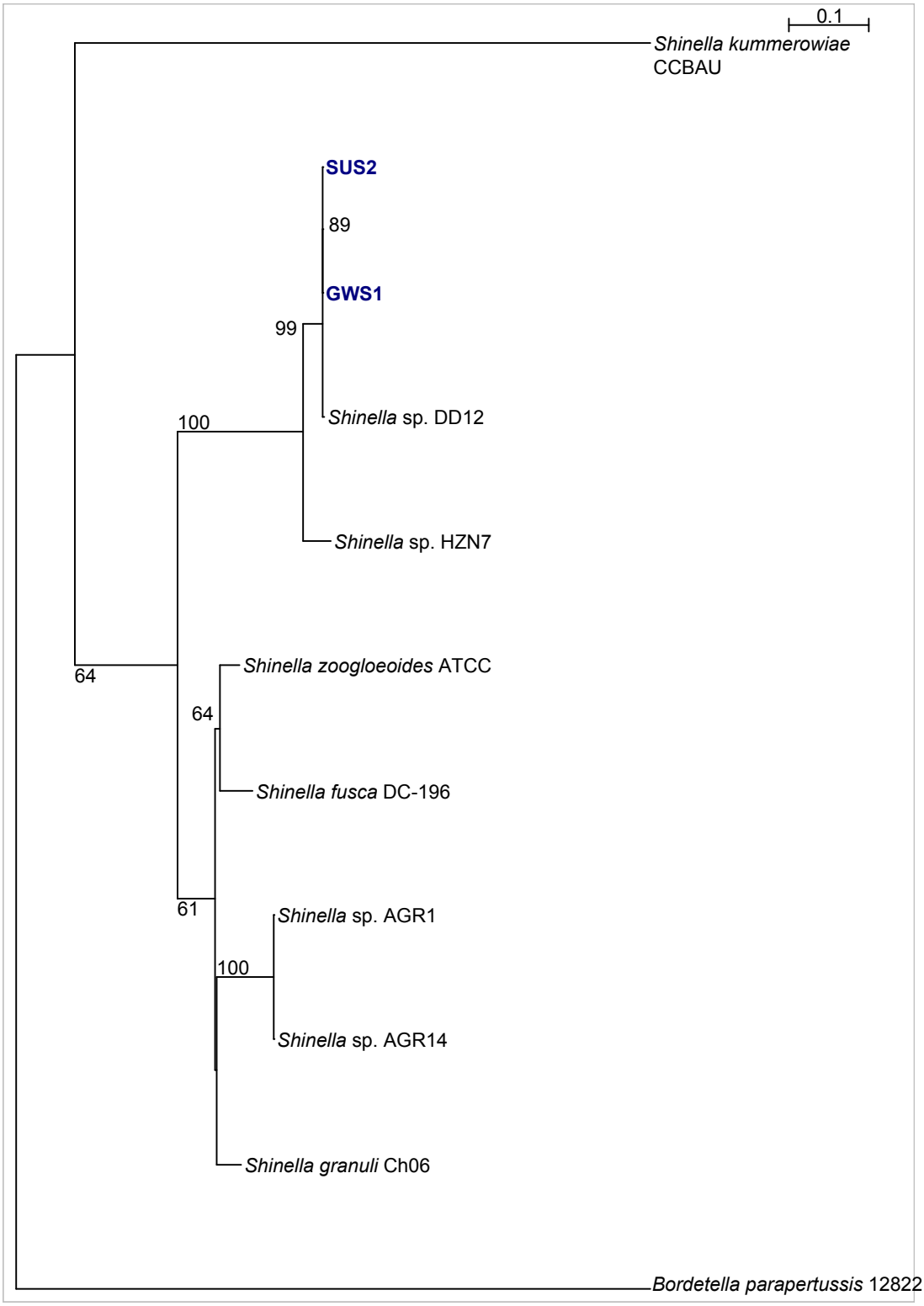


Figure 2.11 Phylogram constructed from maximum likelihood (PhyML) analysis of *atpD* and *recA*, sequence data for bacterial strains GWS1, SUS2 and a range of *Shinella* species. Bootstrapping was set at 100 and values are shown for each node with > 50% bootstrap support. The tree was rooted by placing *Bordetella parapertussis* 12822 as the outgroup.

sp. DD12, isolated from the gut of starved zooplankton *Daphnia magna* (Poehlein *et al.*, 2016). The genome of *Shinella* sp. DD12 encoded three complete pathways for assimilation of phosphonates, indicating a broad ability to be able to utilise reduced phosphonates as P, C and N sources, differentiating this organism from most other *Rhizobiaceae* members. Following on from this, two more *Shinella* genomes were sequenced in 2016: *Shinella* sp. HZN7 and *Shinella* sp. 65-6. *Shinella* sp. HZN7, isolated from the active sludge of a pesticide waste water treatment system in China, was found to have a nicotine-degrading gene cluster present on a plasmid (Qiu *et al.*, 2016). The absence of this gene cluster in other *Shinella* genomes (presumably *Shinella* spp. DD12, SUS2 and GWS1) led the authors to surmise that they may have been acquired as the result of horizontal gene transfer. The genome of *Shinella* sp. 65-6 was assembled from a metagenomic sample obtained from two laboratory-scale bioreactors used for the study of cyanide and thiocyanate degradation (Kantor *et al.*, 2015). The study by Kantor *et al.* did not give any specific information about the genomic features of *Shinella* 65-6 but did suggest that a large portion of the microbial community was autotrophic, gaining energy from the oxidation of sulfur compounds produced during thiocyanate degradation.

Section 2.5.3 will look at genome comparisons between the three *Shinella* genomes previously sequenced and the genomes of *Shinella* spp. GWS1 and SUS2 as well as looking at genomic features which may indicate similar characteristics between *Shinella* sp. SUS2 and GWS1 and other previously discovered *Shinella* species.

2.5.2 *Shinella* strains GWS1 and SUS2 are very closely related

Extraction (using RNAmmer) and alignment (using MUMmer) of the 16S rRNA gene sequence from the assembled genomes of bacterial strains SUS2 and GWS1, showed that the two 16S rRNA sequences were 100% identical. To further determine the sequence similarity between bacterial strains SUS2 and GWS1 their assembled genomes were aligned using Mauve; figure 2.12 shows the Mauve alignment and demonstrates a high level of similarity between the two assembled genomes. Average nucleotide identity (ANI) was calculated at 99.9 % using MUMmer. Additionally, the raw paired-end reads from SUS2 and

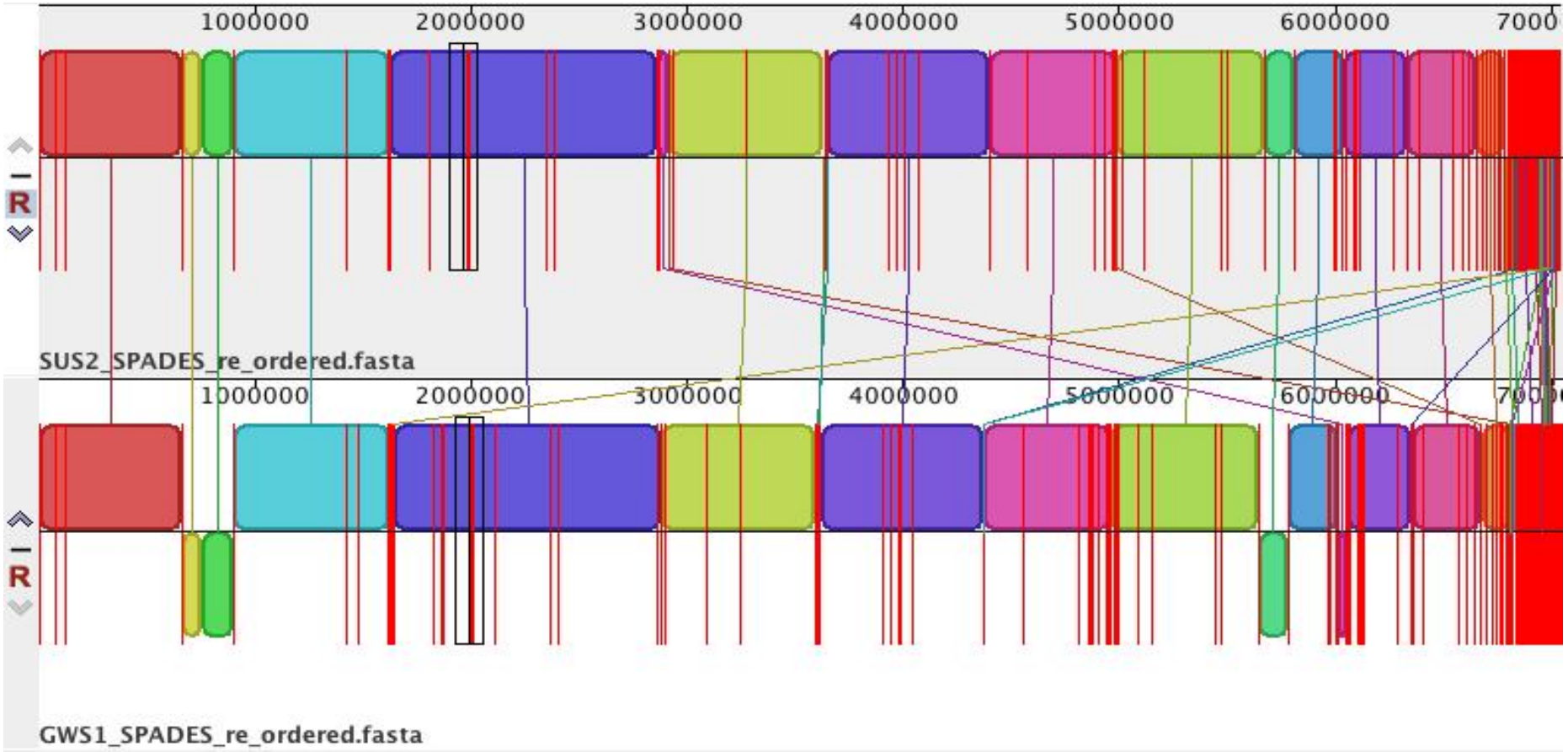


Figure 2.12 Mauve alignment between *Shinella* sp. SUS2 (top) and *Shinella* sp. GWS1 (bottom). Areas of homology between the two sequences are represented with coloured blocks and contig boundaries are represented by a red vertical line. A high level of homology between the two genomes is demonstrated, consistent with their being members of the same species of *Shinella*.

GWS1 were assembled against the *de-novo* assembled genome from the other strain using BWA mem, resulting in 99.72% of SUS2 raw reads mapping to GWS1 and 99.70 of GWS1 raw reads mapping to SUS2. The annotated genomes of *Shinella* sp. SUS2 and GWS1 were compared using Rapid Annotations using Subsystems Technology (RAST) and this found only five genes which were present in SUS2 but not GWS1 encoding: Transcriptional activator of maltose regulon MalT*,* Endo-1,4-beta-xylanase A precursor, 5-methylaminomethyl-2-thiouridine-forming enzyme MnmC*,* ParE toxin protein and uncharacterized monothiol glutaredoxin ycf64-like. Only one gene was present in GWS1 but not SUS2 encoding: Conjugative transfer protein TrbG. Such high levels of similarity between the sequences and gene content of *Shinella* sp. SUS2 and *Shinella* sp. GWS1 indicate they are two strains of the same *Shinella* species.

2.5.3 Whole genome analysis of *Shinella* sp. GWS1 and *Shinella* sp. SUS2

As previously discussed, as of Feb 2017, there were three genomes from the *Shinella* genus available on the NCBI database (https://www.ncbi.nlm.nih.gov/genome/genomes/32494). Figure 2.13 shows BRIG whole genome comparisons between these three genomes and *Shinella* SUS2 (due to their very high levels of similarity, this analysis was carried out on *Shinella* SUS2 only and not *Shinella* GWS1). The whole genome comparison demonstrates that although there are areas of difference in the genomes of *Shinella* sp. SUS2 and the three other strains, the most similar is strain DD12. This is further confirmed with a high ANI score of 98.72% between *Shinella* sp. SUS2 and *Shinella* sp. DD12 indicating they are the same species. Key statistics relating to these genomes are shown in table 2.6. A specific feature of the *Shinella* sp. DD12 genome is that it encodes three complete pathways for assimilation of phosphonates. Additionally, genome analysis of *Shinella* sp. DD12 indicates that the organism is a denitrifier as it has gene encoding two pathways: a dissimilatory nitrate reduction to ammonia pathway and an assimilative nitrate reduction to L-glutamine and L-glutamate pathway (Pohlein *et al.*, 2016). Like all other *Shinella* species so far discovered, with the exception of the symbiotic *S. kummerowia,* genes responsible for nitrogen fixation were absent from *Shinella* DD12. These genes were all looked for in *Shinella* sp. GWS1/SUS2. Table 2.7 shows that the genes involved in the

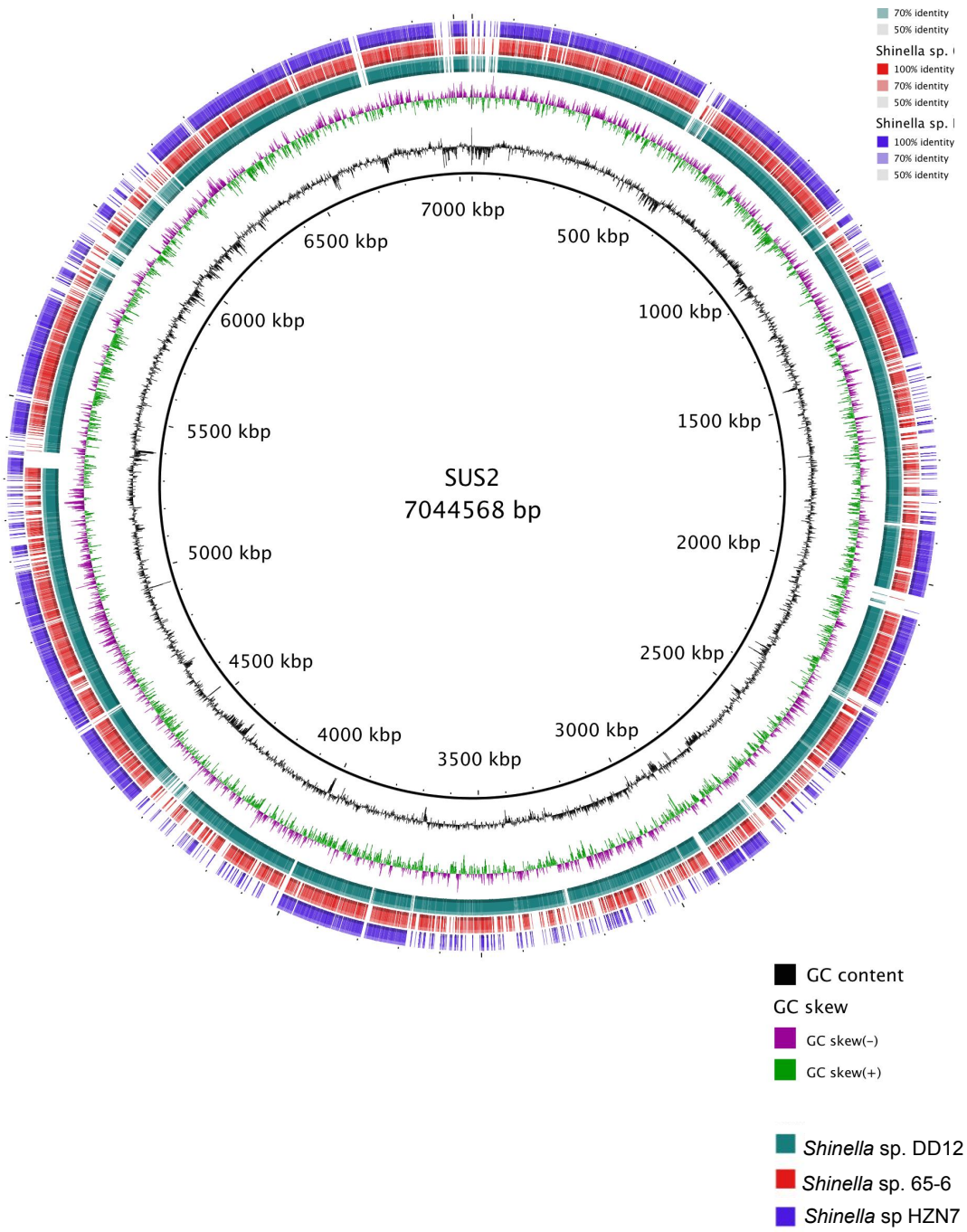


Figure 2.13 Whole genome comparisons between *Shinella* sp. SUS2 and *Shinella* sp. DD12, *Shinella* sp 65-6 and *Shinella* sp. HZN7 created using the BLAST Ring Image Generator (BRIG).

Table 2.6 Whole genome statistics for *Shinella* sp. GWS1, SUS2 and three *Shinella* strains taken from the NCBI genome database.

Shinella strain	Genome size (Mb) *	GC content (%) *	Number of CDS *	ANI to SUS2/GWS1 (%)	Core genome size (% of total genome)	Total number of accessory CDS. **	Predicted phage genes (accessory genome) ***	Number of missing BUSCOs (% of genome)
GWS1	6.98	63.7	6664	-	-	-	-	4 (2.7 %)
SUS2	7.00	63.7	6704	-	68	2297	205	4 (2.7 %)
DD12	7.68	63.4	7393	98.72	62	3035	185	5 (3.3 %)
HZN7	7.35	64.8	6963	91.53	66	2603	162	5 (3.4 %)
65-6 ****	5.12	63.4	5259	87.94	-	-	-	45 (30.5 %)

*According to NCBI
** Total number of protein encoding genes and RNAs (according to RAST) present in accessory gene files generated by Spine
*** Genes contained within intact and incomplete prophage regions predicted by PHASTER
**** *Shinella* sp. 65-6 was not included in core genome analysis due to the smaller than expected genome size and the high number of missing BUSCOs.

Table 2.7 Presence or absence of genes involved in the assimilation of phosphonates, denitrification and nitrogen fixation in *Shinella* spp. GWS1/SUS2 and *Shinella* sp. DD12.

Gene	***Shinella* sp. DD12**	***Shinella* spp. GWS1/SUS2**
Genes involved in the assimilation of phosphonates		
Alkaline phosphatase, *phoA*	+	+
C-P lyase complex, phnGHIKKLM	+	+
Phosphonoaecetaldehyde dehydrogenase, *phnWAY,*	+	+
Genes involved in denitrification		
Periplasmic nitrate reductase, *napABC*	+	+
NO-forming nitrite reductase*, nirK*	+	+
Nitrous oxide reductase *nosZ*	+	+
Genes involved in nitrogen fixation		
nodC	-	-
nifH	-	-

assimilation of phosphonates and genes involved in denitrification are present in *Shinella* sp. GWS1 and SUS2 genomes and nitrogen fixation genes are absent. This adds further weight to the theory that *Shinella* sp. GWS1/SUS2 is the same species as *Shinella* sp. DD12.

A common feature in the genus *Shinella*, as well as many other Rhizobiaceae, is the presence of numerous plasmids (Pohlein *et al.*, 2016; Qiu *et al.*, 2016). Strain DD12 contains at least seven plasmids, detected through the presence of the *repABC* operon in several different locations within the genome (Pohlein *et al.*, 2016). Six *repABC* gene clusters are also present in *Shinella* sp. GWS1 and SUS2 genomes, although *repC* is absent from one cluster in GWS1. The *repABC* operon is composed of plasmid segregation and replication genes, it is unique to Alphaproteobacteria and is a common feature in Rhizobiales (Cavellos *et al*., 2008; Pinto *et al*, 2012.).

The core genome for *Shinella* sp. SUS2, *Shinella* sp. DD12 and *Shinella* sp. HZN7 was constructed using Spine. The genome for *Shinella* sp. 65-6 was not included in core genome analysis due to its distance from *Shinella* sp. SUS2 in the phylogenetic trees as well as its high number of missing BUSCOs (table 2.6). Accessory genomic elements (AGEs) were identified in each of the three genomes using AGEnt and these were then clustered using ClustAGE in order to identify the minimum set of AGEs in the three genomes as well as to determine the distribution of each AGE among the genomes. Results from ClustAGE were plotted using ClustAGE Plot (figure 2.14). Statistics related to the core and accessory genomes are included in table 2.6. FASTA files of accessory genes for each strain were uploaded to RAST for annotation. Figure 2.15 shows subsystem category distribution according to RAST for accessory genes for each *Shinella* strain.

Figure 2.14 shows that *Shinella* sp. SUS2 has a number of accessory genes that are also present in either *Shinella* sp. DD12 or *Shinella* sp. HZN7, with very few genes which are unique to *Shinella* sp. SUS2 alone. Comparisons, using annotations provided by RAST, show that *Shinella* sp. SUS2 has 330 accessory genes which it shares with *Shinella* sp. DD12, and 179 accessory genes which it shares with *Shinella* sp. HZN7. Only 33 genes are present in *Shinella* sp.

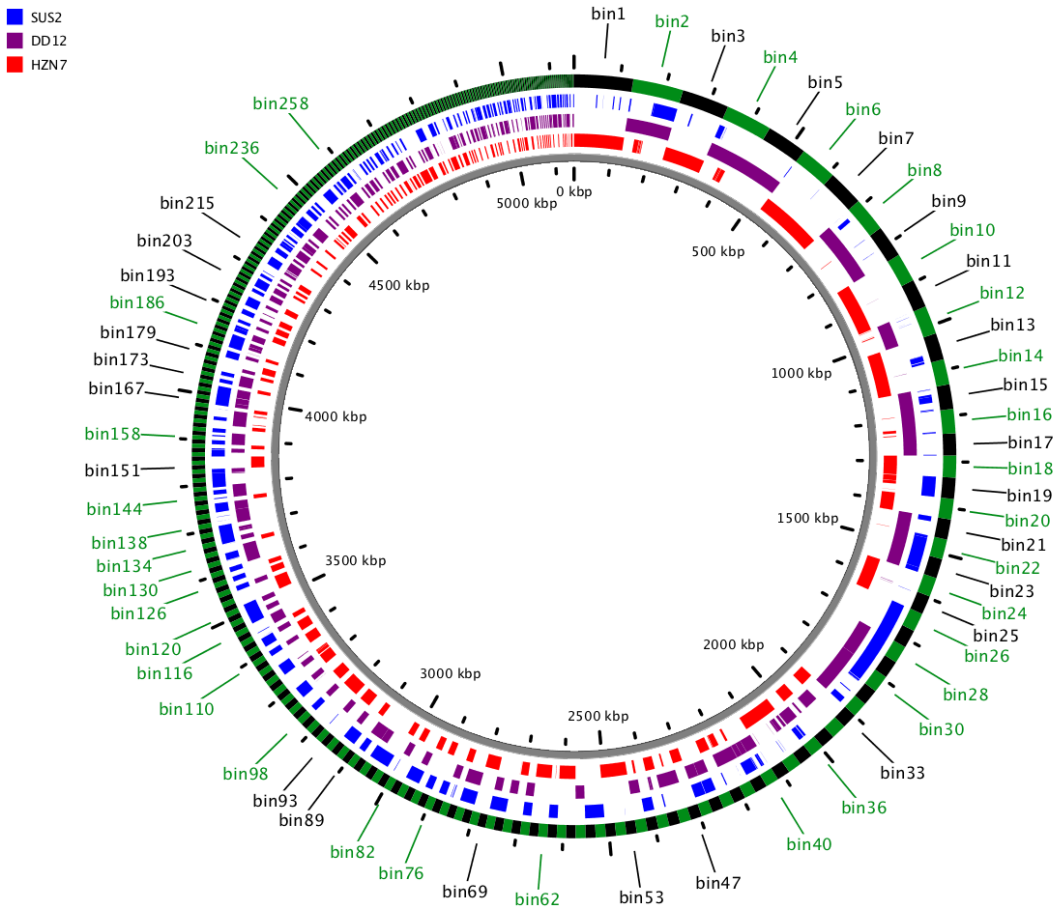


Figure 2.14 Output from Clustage Plot, showing the distribution of accessory genes for *Shinella* sp. SUS2, *Shinella* sp. DD12 and *Shinella* sp. HZN7. Sequences over 5000 bases are labelled (with prefix “bin”).

SUS2 but absent from *Shinella* spp. DD12 and HZN7. These are shown in table 2.8. These comparisons do not include the large numbers of hypothetical proteins present as follows: *Shinella* sp. SUS2, 826; *Shinella* sp. DD12, 999; *Shinella* sp. HZN7, 821.

Figure 2.15 shows that the accessory genome for *Shinella* sp. SUS2 has the highest proportion of genes classified into the RAST subsystems categories of carbohydrates and membrane transport. Included in the membrane transport category are the type VI secretion system (further discussed in section 2.9), cation transporters, tricarboxylate transporters and TRAP (tripartite ATP-independent periplasmic) transporters. Within the carbohydrate category are systems for central carbohydrate metabolism (of methylglyoxal and pyruvate), maltose and maltodextrin utilisation and lactate fermentation. The category of virulence, disease and defence is present in the accessory genome of *Shinella* sp. SUS2; however, these are all genes related to antibiotic resistance, which is most likely due to its prolonged time as part of a lab culture. Biotin synthesis is present, a system of interest due to the link between bacteria and the supply of B vitamins for algae; this is discussed further in section 2.8.

Table 2.8 Genes present in *Shinella* sp. SUS2 but absent from other *Shinella* spp. discussed in this study. Classified by RAST.

Category	**Subcategory**	**Subsystem**	**Role**
Carbohydrates	Di- and oligosaccharides	Maltose and Maltodextrin Utilization	Transcriptional activator of maltose regulon MalT
Carbohydrates	Fermentation	Acetyl-CoA fermentation to Butyrate	3-hydroxybutyryl-CoA dehydratase (EC 4.2.1.55)
Carbohydrates	Fermentation	Acetyl-CoA fermentation to Butyrate	Acetoacetyl-CoA reductase (EC 1.1.1.36)
Carbohydrates	Fermentation	Acetyl-CoA fermentation to Butyrate	Acetyl-CoA acetyltransferase (EC 2.3.1.9)
Carbohydrates	Fermentation	Acetyl-CoA fermentation to Butyrate	Acetyl-CoA:acetoacetyl-CoA transferase, alpha subunit (EC 2.8.3.8)
Carbohydrates	Fermentation	Acetyl-CoA fermentation to Butyrate	Acetyl-CoA:acetoacetyl-CoA transferase, beta subunit (EC 2.8.3.8)
Carbohydrates	Fermentation	Fermentations: Lactate	D-lactate dehydrogenase (EC 1.1.1.28)
Carbohydrates	Fermentation	Fermentations: Lactate	Phosphate acetyltransferase (EC 2.3.1.8)
Carbohydrates	One-carbon Metabolism	One-carbon metabolism by tetrahydropterines	Methenyltetrahydrofolate cyclohydrolase (EC 2.5.4.9)
Carbohydrates	One-carbon Metabolism	One-carbon metabolism by tetrahydropterines	Methylenetetrahydrofolate dehydrogenase (NADP+) (EC 1.5.1.5)
Carbohydrates	Polysaccharides	Glycogen metabolism	Glycogen phosphorylase (EC 2.4.1.1)
Clustering-based subsystems	DNA polymerase III epsilon cluster	CBSS-342610.3.peg.1536	Membrane-bound lytic murein transglycosylase D precursor (EC 2.2.1.-)
Clustering-based subsystems	no subcategory	Aminoglycoside phosphotransferase (antibiotic) cluster	Aminoglycoside 3'-phosphotransferase 2 (EC 2.7.1.95)
Clustering-based subsystems	no subcategory	CBSS-374931.9.peg.1048	FIG001353: Acetyltransferase
Cofactors, Vitamins, Prosthetic Groups	Biotin	Biotin biosynthesis	3-ketoacyl-CoA thiolase (EC 2.3.1.16)
DNA Metabolism	DNA replication	DNA replication strays	Error-prone repair homolog of DNA polymerase III alpha subunit (EC 2.7.7.7)
DNA Metabolism	DNA replication	Plasmid replication	Chromosome (plasmid) partitioning protein ParB
DNA Metabolism	no subcategory	Restriction-Modification System	Type III restriction-modification system

			methylation subunit (EC 2.1.1.72)
Fatty Acids, Lipids, and Isoprenoids	no subcategory	Polyhydroxybutyrate metabolism	Polyhydroxyalkanoic acid synthase
Membrane Transport	ABC transporters	ABC transporter dipeptide (TC 2.A.1.5.2)	Dipeptide transport ATP-binding protein DppF (TC 2.A.1.5.2)
Miscellaneous	no subcategory	Muconate lactonizing enzyme family	L-rhamnonate dehydratase (EC 4.2.1.90)
Phages, Prophages, Transposable elements, Plasmids	Phages, Prophages	Phage lysis modules	Phage endolysin
Phages, Prophages, Transposable elements, Plasmids	Phages, Prophages	Phage replication	DNA transposition protein
Regulation and Cell signaling	Programmed Cell Death & Toxin-antitoxin Systems	Phd-Doc	YdcE-YdcD toxin-antitoxin systems FIG022160: hypothetical toxin
Regulation and Cell signaling	Programmed Cell Death & Toxin-antitoxin Systems	Phd-Doc	YdcE-YdcD toxin-antitoxin systems FIG045511: hypothetical antitoxin (to FIG022160: hypothetical toxin)
Regulation and Cell signaling	no subcategory	DNA-binding regulatory proteins, strays	LysR family transcriptional regulator PA2877
Stress Response	Osmotic stress	Choline and Betaine Uptake and Betaine Biosynthesis	High-affinity choline uptake protein BetT
Stress Response	Oxidative stress	Glutathione: Non-redox reactions	Glutathione S-transferase family protein
Stress Response	Oxidative stress	Oxidative stress	Peroxidase (EC 1.11.1.7)
Stress Response	Oxidative stress	Oxidative stress	transcriptional regulator, Crp/Fnr family
Virulence, Disease and Defense	Resistance to antibiotics and toxic compounds	Cobalt-zinc-cadmium resistance	Cation efflux system protein CusA
Virulence, Disease and Defense	Resistance to antibiotics and toxic compounds	Cobalt-zinc-cadmium resistance	Cobalt-zinc-cadmium resistance protein CzcA
Virulence, Disease and Defense	Resistance to antibiotics and toxic compounds	Copper homeostasis	Copper chaperone

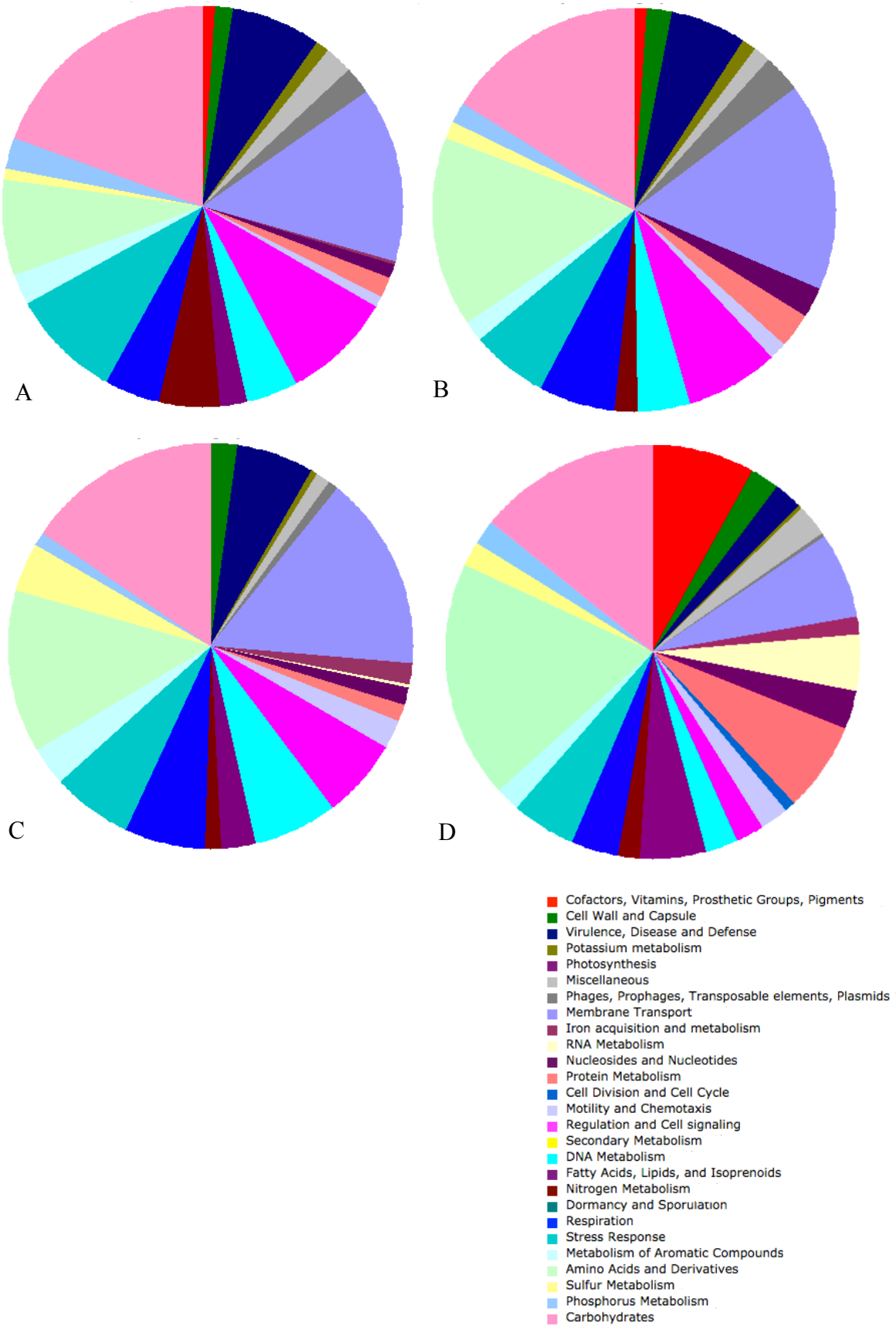


Figure 2.15. Accessory gene classification according to RAST, for A = *Shinella* sp. SUS2, B = *Shinella* sp. DD12, C = *Shinella* sp. HZN7. Figure D shows the core genome

2.6 Bacterial strain SUL3 is a member of the genus *Agrobacterium.*

2.6.1 Phylogenetic analysis of bacterial strain SUL3

In order to determine the taxonomy of bacterial strain SUL3 the 16S rRNA gene sequence was extracted from its assembled genome. This was used as the query in a BLAST search of the NCBI 16s rRNA (bacteria and archaea) database and all sequences were downloaded which had at least 95% identity to bacterial strain SUL3. A maximum likelihood phylogenetic tree was constructed using the 16S rRNA gene sequences (Figure 2.16). The 16S rRNA phylogenetic tree places SUL3 in a clade with a range of mostly *Agrobacterium* and *Rhizobium*. *Beijerinckia fluminensis* UQM, which is close to SUL3, has been reclassified as *Rhizobium radiobacter*, which is the updated scientific name for *Agrobacterium tumefaciens* (Oggerin *et al.*, 2009). The housekeeping genes *recA* and *atpD* were extracted from the genomes of bacterial strain SUL3 and from the genomes of a representation of bacteria from the Rhizobiaceae family (taken from the NCBI database) which were present in the 16S rRNA phylogenetic tree. These *recA* and *atpD* gene sequences were then used to construct a maximum likelihood phylogenetic tree (Figure 2.17). *recA* and *atpD* were chosen for this as they had previously been used in a study by Gaunt *et al*. (2001), along with 16S rRNA, for phylogenetic analysis of Alpha-proteobacteria, including *Agrobacterium*. The *recA* and *atpD* tree also places SUL3 in a clade with predominantly *Rhizobium radiobacter* (formerly *Agrobacterium tumefaciens)*.

It is worth noting the changeable nature of bacterial classification and the complicated history of *Agrobacterium* and *Rhizobium,* both members of the family Rhizobiaceae. With the advent of DNA sequencing it became apparent that the two genera could not easily be distinguished from one another (Willems and Collins, 1993). Young *et al.* (2001) proposed the reclassification of *Agrobacterium* making it a synonym of *Rhizobium.* However, Farrand *et al*.,

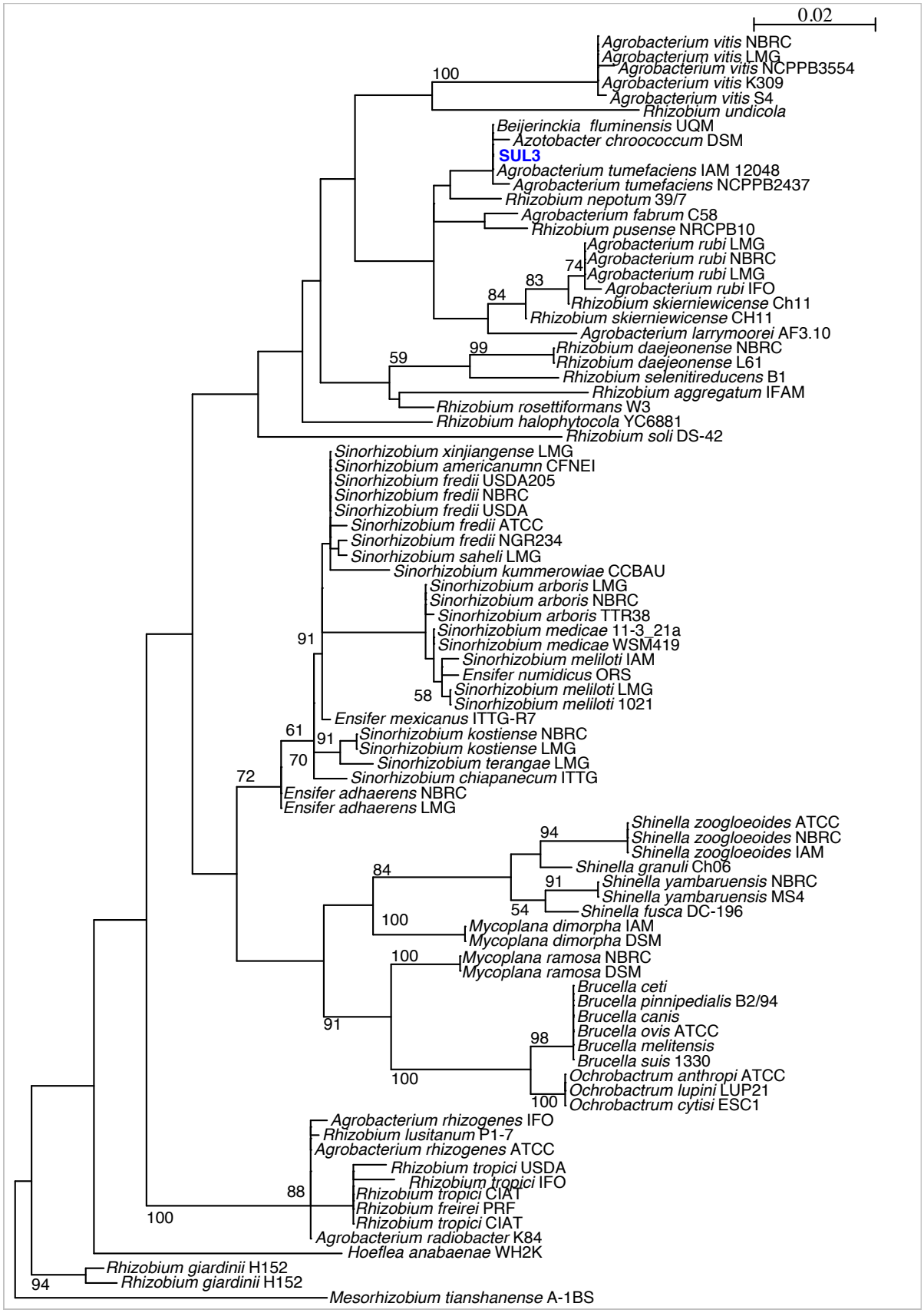


Figure 2.16 Phylogram constructed from maximum likelihood analysis (PhyML) of 16S rRNA gene sequence data bacteria, identified through BLAST as having > 95% identity to 16S rRNA genes of bacterial strain SUL3. Bootstrapping was set at 100 and values are shown for each node with > 50% bootstrap support. The tree is rooted with *Mesorhizobium*

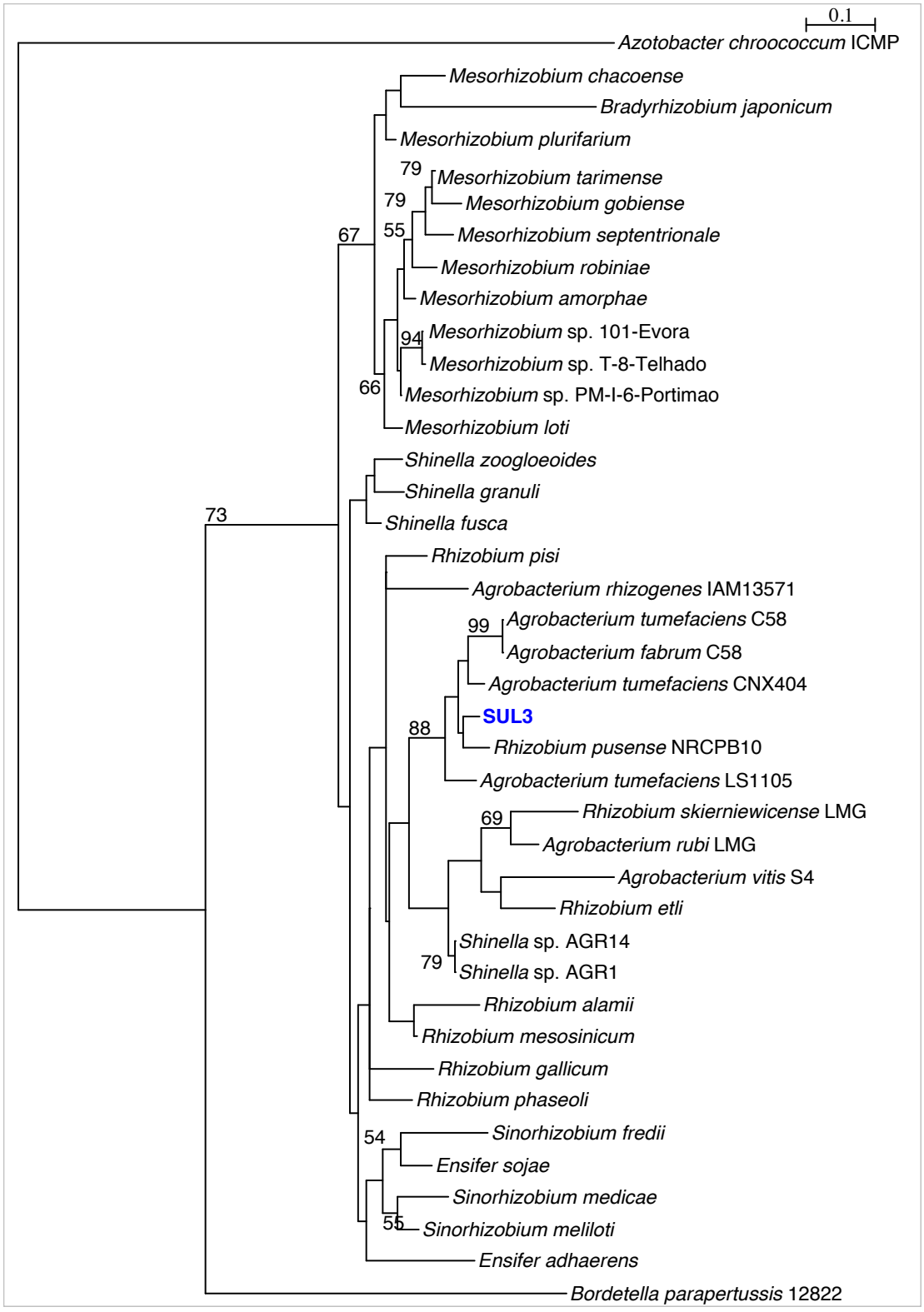


Figure 2.17 Phylogram constructed from maximum likelihood analysis (PhyML) of *atpD* and *recA* sequence data for bacterial strain SUL3 and bacteria from the Rhizobiaceae family. Bootstrapping was set at 100 and values are shown for each node with > 50% bootstrap support. The tree is rooted with *Bordetella parapertussis* 12822.

(2002) contested this change in classification, claiming the different phenotypic traits between *Agrobacterium* and *Rhizobium* justified the retention of the *Agrobacterium* classification. Despite these protests the reclassification was widely accepted and a large proportion of taxonomic publications now use *Rhizobium* over *Agrobacterium* though the debate is not fully resolved and *Agrobacterium* is still widely used by some scientists (Kuzmanović *et al*., 2015).

2.6.2 *Agrobacterium* sp. SUL3 is the same species as a strain of *Agrobacterium* isolated from an oligotrophic site.

Since submitting *Agrobacterium* sp. SUL3 genome sequence to the GenBank and carrying out the previous phylogenetic analysis (August, 2016), two further similar genomes have also been deposited in the NCBI databases: *Agrobacterium* sp. LC34 and *Rhizobium* sp. Root651 have an ANI to *Agrobacterium* sp. SUL3 of 98.31 % and 98.21 % respectively. This high ANI indicates they are members of the same species. Figures 2.18 and 2.19 show revised 16S and *recA*/*atpD* phylogenetic trees (respectively) including these new strains. Figure 2.20 shows whole genome comparisons between *Agrobacterium* sp. SUL3, *Agrobacterium* sp. LC34 and *Rhizobium* sp. Root 651. Key statistics regarding these genomes are shown in table 2.9

Agrobacterium sp. LC34 was isolated from a deep subsurface (-400 m), oligotrophic site in Lechuguilla Cave, New Mexico, USA. The only available literature regarding *Agrobacterium* sp. LC34 is a conference abstract (Gan *et al*., 2016) which compares this strain to *Agrobacterium* sp. SUL3 and *Rhizobium* sp. Root 651, concluding that they are members of the same genospecies, with strain LC34 especially well adapted to living in low nutrient environments. *Rhizobium* sp. Root 651 was one of 400 isolates taken from the microbiota of *Arabidopsis thaliana*, which were then further classified depending on whether they came from the leaf, root or soil (Bai *et al,* 2015). In order to look for genes that are unique as well as features that are shared the core genome was computed for *Agrobacterium* sp. SUL3, *Agrobacterium* sp. LC34 and *Rhizboum* sp. Root 651 (Using Spine); from this accessory genes were determined in each strain (using AGEnt) and plotted using ClustAGE plot (figure 2.21). Statistics are shown in table 2.9.

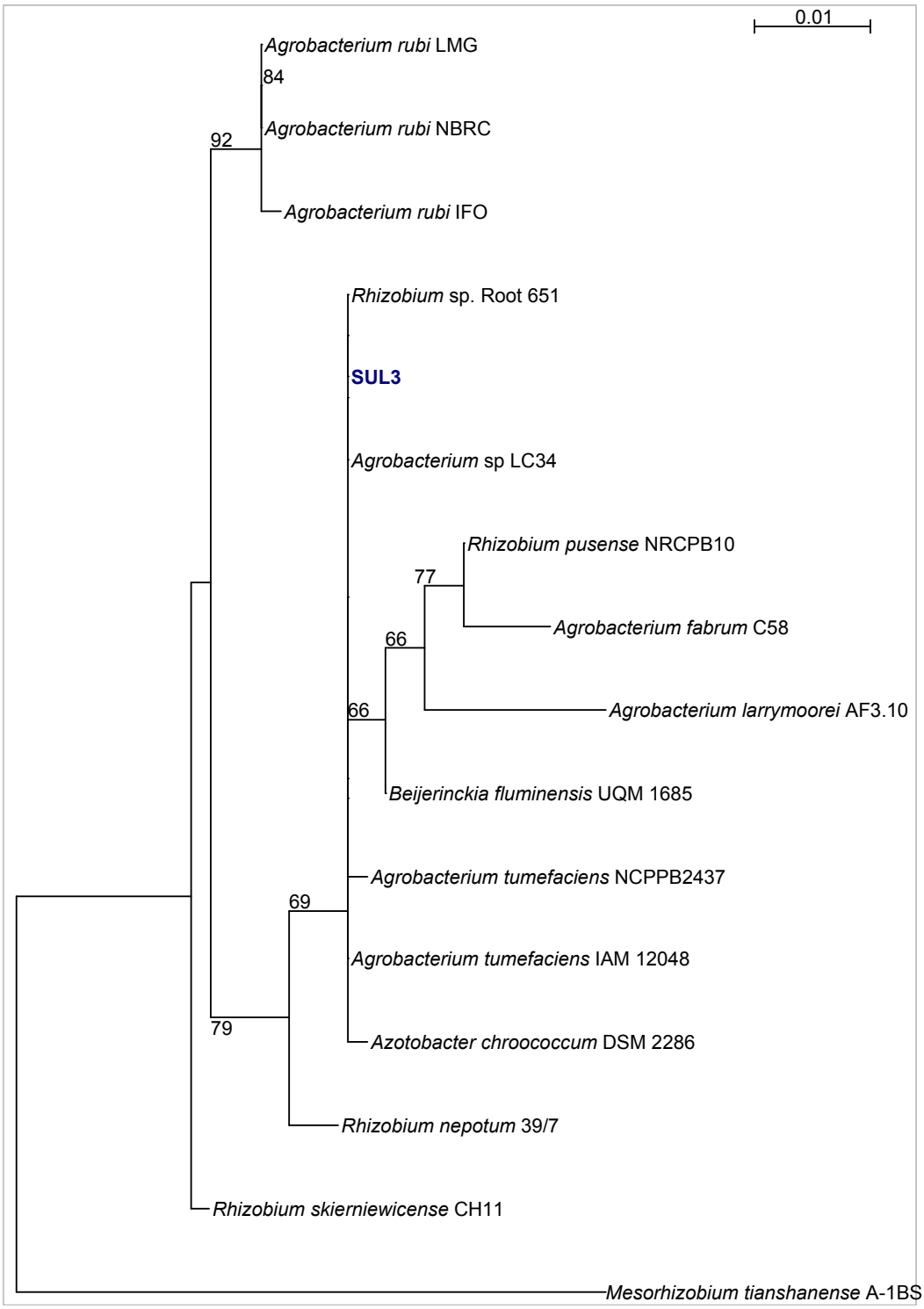


Figure 2.18 Phylogram constructed from maximum likelihood analysis (PhyML) of 16S rRNA gene sequence data for bacterial strain SUL3, bacteria identified from figure 2.16 as phylogenetically close to bacterial strain SUL3, *Rhizobium* sp. Root 651 and *Agrobacterium* sp. LC34. Bootstrapping was set at 100 and values are shown for each node with > 50% bootstrap support. The tree is rooted with *Mesorhizobium tianshanense* A-1BS as outgroup.

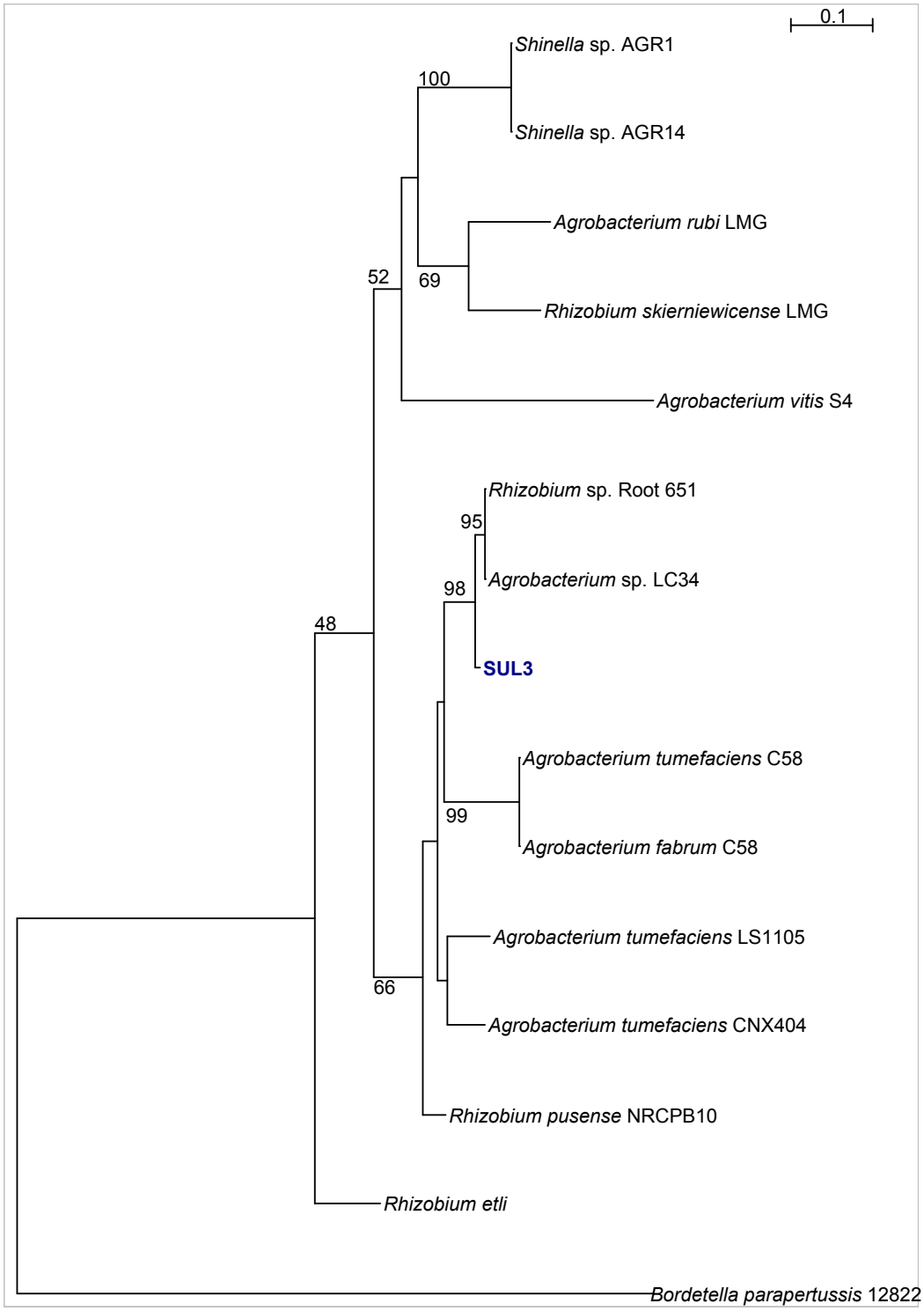


Figure 2.19 Phylogram constructed from maximum likelihood analysis (PhyML) of *atpD* and *recA* sequence data for bacterial strain SUL3, bacteria identified from figure 2.17 as phylogenetically close to bacterial strain SUL3, *Rhizobium* sp. Root 651 and *Agrobacterium* sp. LC34.. Bootstrapping was set at 100 and values are shown for each node with > 50% bootstrap support. The tree is rooted with *Bordetella parapertussis* 12822.

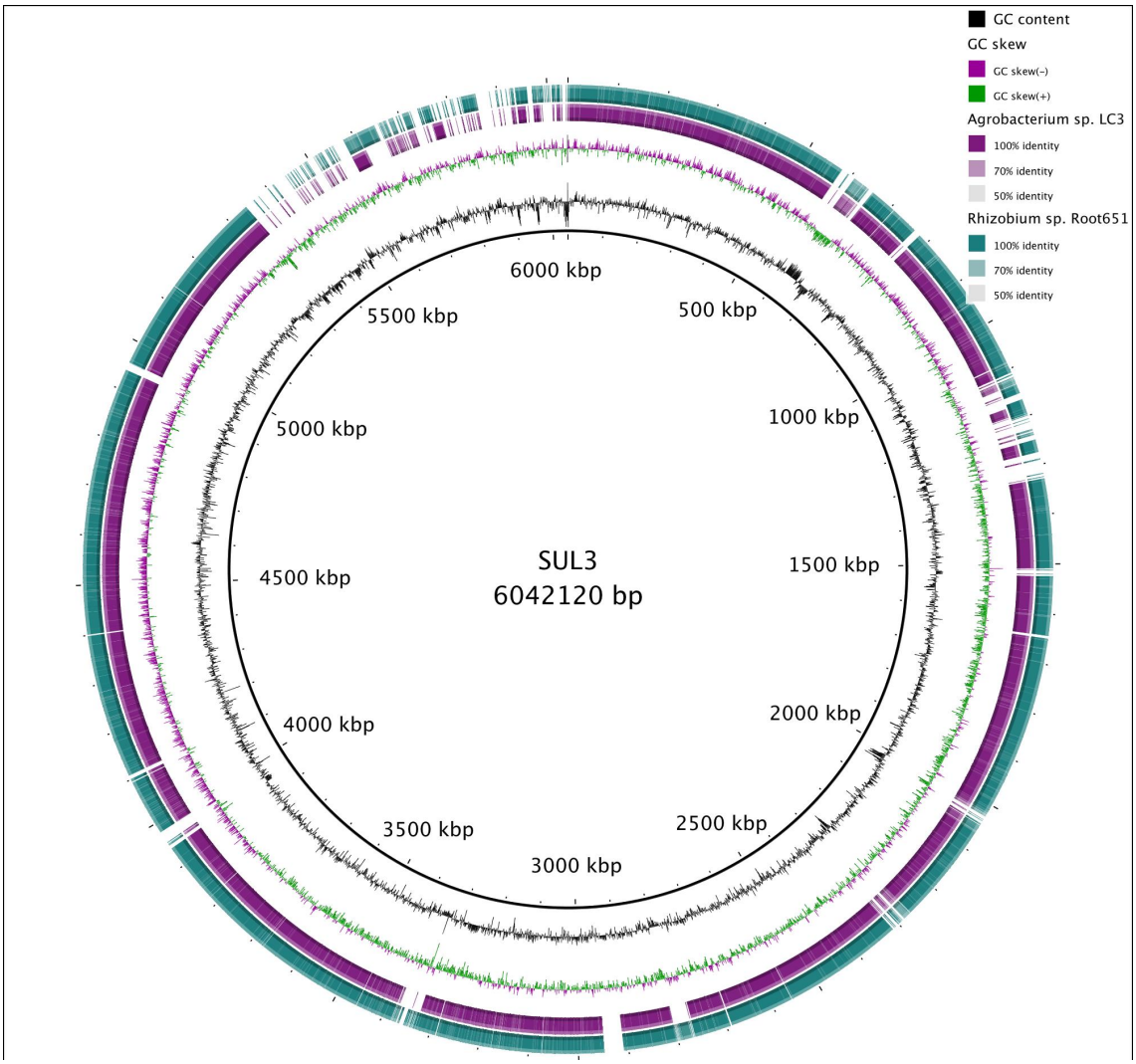


Figure 2.20 Whole genome comparisons between *A. tumefaciens* SUL3 (reference genome), *Agrobacterium* sp. LC34 and *Rhizobium* sp Root 651. created using the BLAST Ring Image Generator (BRIG).

Table 2.9 Genome statistics for *Agrobacterium* sp. SUL3, *Agrobacterium* sp. LC34 and *Rhizobium* sp. Root 652

Bacterial strain	Genome size (Mb) *	GC content (%) *	CDS *	ANI to SUL3 (%)	Core genome size (% of total genome)	Total number of accessory CDS.**	Predicted phage genes in accessory genome ***	Number of missing BUSCOs (% of genome)
Agrobacterium sp. SUL3	6.1	59.2	5802	-	79	1455	172	6 (4 %)
Agrobacterium sp. LC34	5.6	59.4	5257	98.31	88	714	50	6 (4 %)
Rhizobium sp. Root 651	5.8	59.4	5550	98.21	82	1178	229	6 (4 %)

* According to NCBI
** Total number of protein encoding genes and RNAs (according to RAST) present in accessory gene files generated by Spine
*** Genes contained within intact and incomplete prophage regions predicted by PHASTER

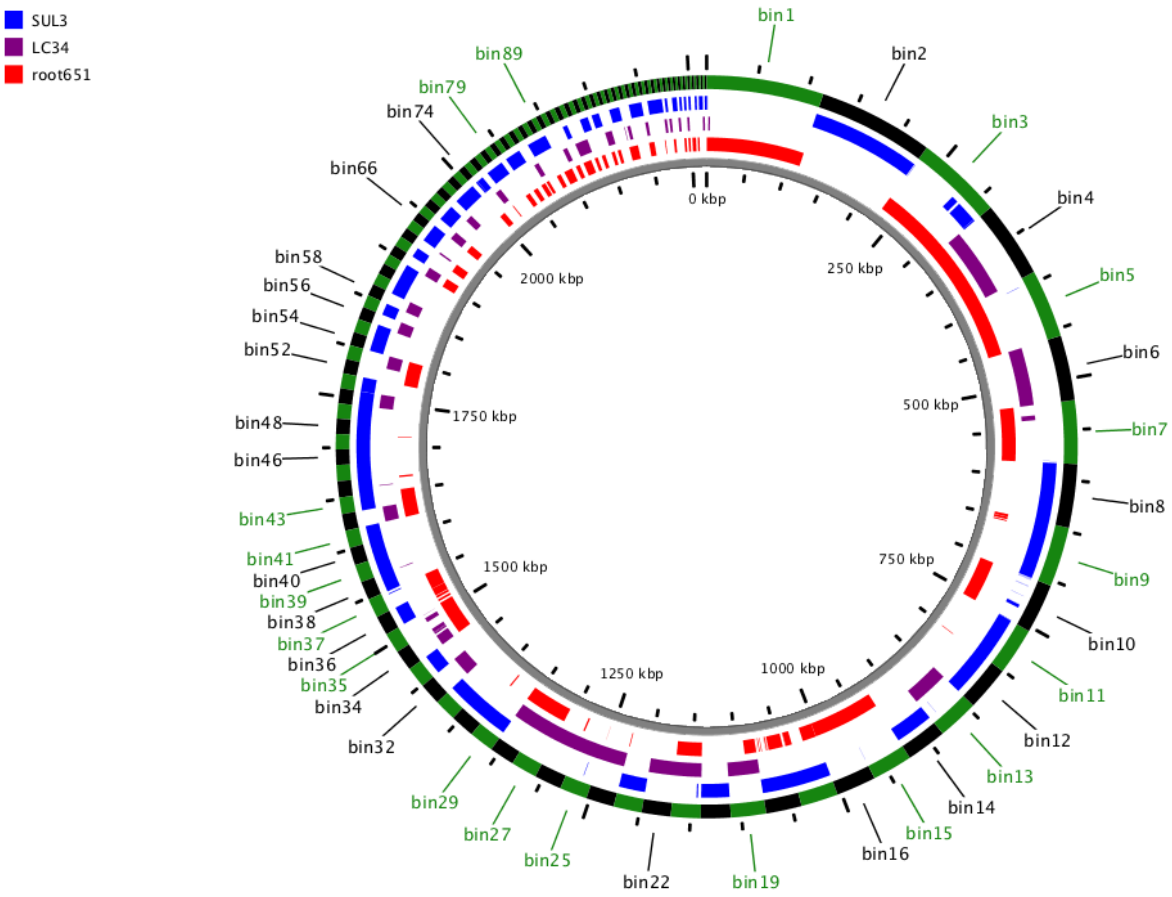


Figure 2.21 Output from Clustage Plot, showing the distribution of accessory genes for *Agrobacterium* sp. SUL3, *Agrobacterium* sp. LC34 and *Rhizobium* sp. Root 651. Sequences over 5000 bases are labelled (with prefix “bin”).

Figure 2.22 shows that the accessory genomes for *Agrobacterium* sp. SUL3, *Agrobacterium* sp. LC34 and *Rhizobium* sp. Root 651 have very different distribution of RAST categories compared to the accessory genomes of the *Shinella* species (Section 2.5.3) with large numbers of genes involved in membrane transport. Within *Agrobacterium* sp. SUL3 the majority of genes involved in this category are conjugative transfer proteins, as discussed in section 2.6.3, below. Conjugative transfer proteins are involved in horizontal gene transfer and may be playing a role in *Agrobacterium* sp. SUL3 interacting with *B. braunii.* Genes present in *Agrobacterium* sp. SUL3 but absent from both *Agrobacterium* sp. LC34 and *Rhizobium* sp. Root 651 are shown in table 2.10.

2.6.3 Plasmid analysis of *Agrobacterium* sp. SUL3

Phylogenetic and whole genome analysis demonstrates that bacterial strain SUL3 is a strain of *Agrobacterium.* A typical characteristic of many *Agrobacterium* species are their pathogenic interactions with plants (Matveeva & Lutova, 2014). A key component in the ability of *Agrobacterium tumefaciens* to induce tumours in plants is the presence of a tumour-inducing (Ti) plasmid (Van Larebeke *et al*., 1974). In order to determine if a Ti plasmid is present in *Agrobacterium* sp. SUL3, BLAST was used with the genome of *Agrobacterium* sp. SUL3 as the database and a Ti plasmid sequence as the query. This found no significant hits. RAST annotations also found no evidence of a plasmid in *Agrobacterium* sp. SUL3.

An alignment, using Mauve, between the genome sequence of *A. tumefaciens* SUL3 and the whole genome sequence of *A. fabrum* C58 (also known as *Agrobacterium tumefaciens* C58), including both the Ti and At plasmids, was carried out and shows a small area of identity between *Agrobacterium* sp. SUL3 and the Ti plasmid region of *A. fabrum* C58 (Figure 2.23). This region of identity is located on scaffold 21 (78921 base pairs) of *Agrobacterium* sp. SUL3. Scaffold 21 was extracted and used as the query in a BLAST search against the NCBI database. The results from this BLAST search contained sequences from a number of plasmids from mostly *Rhizobium* and *Agrobacterium* (Table 2.11). *Rhizobium* and *Agrobacterium* are both members of the Rhizobiaceae family, members of which are able to form commensal or pathogenic relationships with plants (Yanagi and Yamasato, 1993). The nature of this

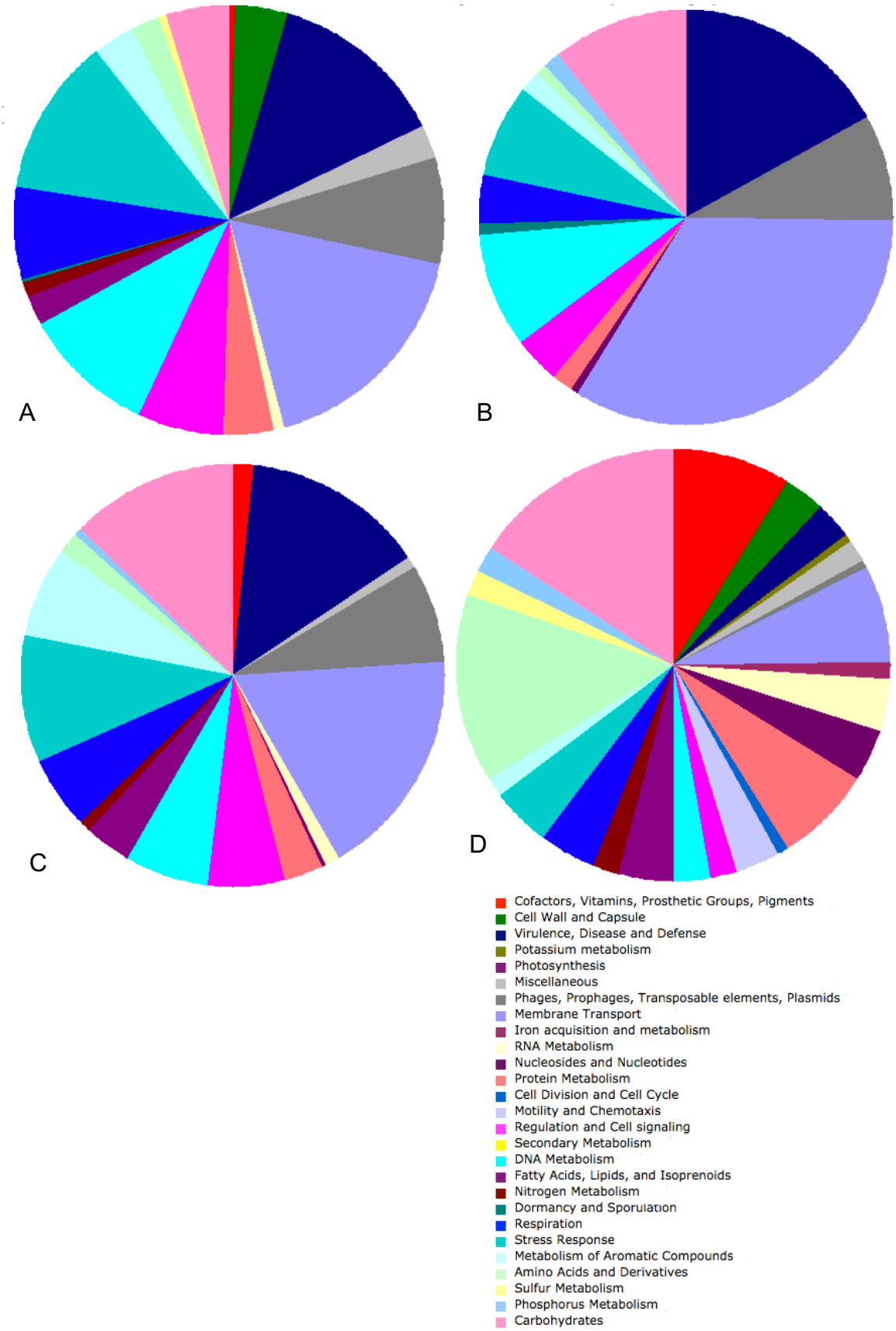


Figure 2.22 Accessory gene classification according to RAST, for A = *Agrobacterium* sp. SUL3, B = *Agrobacterium* sp. LC34, C = *Rhizobium* sp. Root 65. Figure D shows the core genome.

Table 2.10 Genes unique to *Agrobacterium* sp. SUL3. classified according to RAST.

Category	**Subcategory**	**Subsystem**	**Role**
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Polyamine Metabolism	ABC transporter, periplasmic spermidine putrescine-binding protein *PotD* (TC 2.A.1.11.1)
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Polyamine Metabolism	Agmatinase (EC 2.5.3.11)
Amino Acids and Derivatives	Glutamine, glutamate, aspartate, asparagine; ammonia assimilation	Glutamine, Glutamate, Aspartate and Asparagine Biosynthesis	Aspartate aminotransferase (EC 2.6.1.1)
Amino Acids and Derivatives	no subcategory	Creatine and Creatinine Degradation	Creatinine amidohydrolase (EC 2.5.2.10)
Carbohydrates	Aminosugars	Chitin and N-acetylglucosamine utilization	N-Acetyl-D-glucosamine ABC transport system, sugar-binding protein
Carbohydrates	Central carbohydrate metabolism	Dihydroxyacetone kinases	Dihydroxyacetone kinase, ATP-dependent (EC 2.7.1.29)
Carbohydrates	Fermentation	Acetyl-CoA fermentation to Butyrate	3-hydroxyacyl-CoA dehydrogenase (EC 1.1.1.35)
Carbohydrates	Fermentation	Acetyl-CoA fermentation to Butyrate	3-hydroxybutyryl-CoA dehydrogenase (EC 1.1.1.157)
Carbohydrates	Fermentation	Acetyl-CoA fermentation to Butyrate	Acetoacetyl-CoA reductase (EC 1.1.1.36)
Carbohydrates	Monosaccharides	Mannose Metabolism	Mannose-6-phosphate isomerase (EC 5.3.1.8)
Cell Wall and Capsule	Capsular and extracellular polysacchrides	Rhamnose containing glycans	Capsular polysaccharide biosynthesis/export periplasmic protein *WcbA*
Cell Wall and Capsule	Capsular and extracellular polysacchrides	Rhamnose containing glycans	Capsular polysaccharide export system protein *KpsC*

Cell Wall and Capsule	Capsular and extracellular polysaccharides	Rhamnose containing glycans	Glucose-1-phosphate thymidylyltransferase (EC 2.7.7.24)
Cell Wall and Capsule	Capsular and extracellular polysaccharides	Rhamnose containing glycans	dTDP-4-dehydrorhamnose 3,5-epimerase (EC 5.1.3.13)
Cell Wall and Capsule	Capsular and extracellular polysaccharides	Rhamnose containing glycans	dTDP-4-dehydrorhamnose reductase (EC 1.1.1.133)
Cell Wall and Capsule	Capsular and extracellular polysaccharides	Rhamnose containing glycans	dTDP-glucose 4,6-dehydratase (EC 4.2.1.46)
Cell Wall and Capsule	no subcategory	Murein Hydrolases	Membrane-bound lytic murein transglycosylase C precursor (EC 2.2.1.-)
Cofactors, Vitamins, Prosthetic Groups, Pigments	Riboflavin, FMN, FAD	Flavodoxin	Flavodoxin 2
Cofactors, Vitamins, Prosthetic Groups, Pigments	Riboflavin, FMN, FAD	Flavodoxin	NAD(P)H oxidoreductase YRKL (EC 1.6.99.-)
DNA Metabolism	DNA repair	DNA repair, bacterial UvrD and related helicases	DNA helicase IV
DNA Metabolism	no subcategory	DNA ligases	ATP-dependent DNA ligase (EC 6.5.1.1) LigC
DNA Metabolism	no subcategory	Restriction-Modification System	Type I restriction-modification system, DNA-methyltransferase subunit M (EC 2.1.1.72)
DNA Metabolism	no subcategory	Restriction-Modification System	Type I restriction-modification system, restriction subunit R (EC 2.1.21.3)
DNA Metabolism	no subcategory	Restriction-Modification System	Type I restriction-modification system, specificity subunit S (EC 2.1.21.3)
Membrane Transport	ABC transporters	ABC transporter dipeptide (TC 2.A.1.5.2)	Dipeptide transport system permease protein DppB (TC 2.A.1.5.2)

Membrane Transport	ABC transporters	ABC transporter dipeptide (TC 2.A.1.5.2)	Dipeptide transport system permease protein DppC (TC 2.A.1.5.2)
Membrane Transport	ABC transporters	ABC transporter dipeptide (TC 2.A.1.5.2)	Dipeptide-binding ABC transporter, periplasmic substrate-binding component (TC 2.A.1.5.2)
Membrane Transport	ABC transporters	ABC transporter oligopeptide (TC 2.A.1.5.1)	Oligopeptide transport system permease protein OppB (TC 2.A.1.5.1)
Membrane Transport	Cation transporters	Copper Transport System	Repressor CsoR of the *copZA* operon
Membrane Transport	Cation transporters	Copper transport and blue copper proteins	Pseudoazurin
Membrane Transport	Cation transporters	Magnesium transport	Magnesium and cobalt transport protein *CorA*
Membrane Transport	Cation transporters	Transport of Nickel and Cobalt	Cobalt ABC transporter, permease component CbtK
Membrane Transport	Cation transporters	Transport of Nickel and Cobalt	Predicted cobalt transporter CbtA
Membrane Transport	Protein and nucleoprotein secretion system, Type IV	Conjugative transfer	Ync
Membrane Transport	Protein and nucleoprotein secretion system, Type IV	Conjugative transfer	Ynd
Membrane Transport	TRAP transporters	TRAP Transporter collection	TRAP-type C4-dicarboxylate transport system, large permease component
Metabolism of Aromatic Compounds	Metabolism of central aromatic intermediates	Homogentisate pathway of aromatic compound degradation	Fumarylacetoacetase (EC 2.7.1.2)
Metabolism of Aromatic Compounds	Metabolism of central aromatic intermediates	Homogentisate pathway of aromatic compound degradation	Homogentisate 1,2-dioxygenase (EC 1.13.11.5)
Metabolism of Aromatic Compounds	no subcategory	Gentisate degradation	Maleylacetoacetate isomerase (EC 5.2.1.2)
Miscellaneous	no subcategory	Broadly distributed proteins not in subsystems	UPF0028 protein YchK
Miscellaneous	no subcategory	Inner membrane proteins	DedA protein

Nitrogen Metabolism	no subcategory	Nitrate and nitrite ammonification	Cytochrome c-type protein *NapC*
Nitrogen Metabolism	no subcategory	Nitrosative stress	Nitrite-sensitive transcriptional repressor *NsrR*
Phages, Prophages, Transposable elements, Plasmids	Phages, Prophages	Phage replication	DNA polymerase III alpha subunit (EC 2.7.7.7)
Phages, Prophages, Transposable elements, Plasmids	Phages, Prophages	Phage replication	DNA transposition protein
Phages, Prophages, Transposable elements, Plasmids	Phages, Prophages	Phage replication	Phage DNA replication protein
Phages, Prophages, Transposable elements, Plasmids	Phages, Prophages	Phage tail fiber proteins	Phage tail fibers
Phages, Prophages, Transposable elements, Plasmids	Phages, Prophages	Phage tail proteins	Phage tail protein
Protein Metabolism	Protein degradation	Metallocarboxypeptidases (EC 2.4.17.-)	D-alanyl-D-alanine carboxypeptidase (EC 2.4.16.4)
Protein Metabolism	Protein folding	Peptidyl-prolyl cis-trans isomerase	Survival protein SurA precursor (EC 5.2.1.8)
Protein Metabolism	Protein processing and modification	Peptide methionine sulfoxide reductase	Peptide methionine sulfoxide reductase MsrA (EC 1.8.4.11)
Protein Metabolism	Protein processing and modification	Peptide methionine sulfoxide reductase	Peptide methionine sulfoxide reductase MsrB (EC 1.8.4.12)
Protein Metabolism	Protein processing and modification	Signal peptidase	Signal peptidase I (EC 2.4.21.89)
RNA Metabolism	Transcription	Transcription factors bacterial	Transcription termination protein NusA
Regulation and Cell signaling	Programmed Cell Death and Toxin-antitoxin Systems	Phd-Doc, YdcE-YdcD toxin-antitoxin (programmed cell death) systems	FIG022160: hypothetical toxin
Regulation and Cell signaling	Programmed Cell Death and Toxin-antitoxin Systems	Phd-Doc, YdcE-YdcD toxin-antitoxin (programmed cell death) systems	FIG045511: hypothetical antitoxin (to FIG022160: hypothetical toxin)
Regulation and Cell signaling	Programmed Cell Death	Toxin-antitoxin replicon	ParD protein (antitoxin to ParE)

	and Toxin-antitoxin Systems	stabilization systems	
Regulation and Cell signaling	no subcategory	LysR-family proteins in Escherichia coli	Chromosome initiation inhibitor
Regulation and Cell signaling	no subcategory	LysR-family proteins in Escherichia coli	*cyn* operon transcriptional activator
Regulation and Cell signaling	no subcategory	LysR-family proteins in Escherichia coli	LysR family transcriptional regulator PerR
Regulation and Cell signaling	no subcategory	LysR-family proteins in Escherichia coli	*LysR* family transcriptional regulator YnfL
Regulation and Cell signaling	no subcategory	Orphan regulatory proteins	Glycine cleavage system transcriptional activator
Regulation and Cell signaling	no subcategory	Orphan regulatory proteins	Sensory histidine kinase *QseC*
Regulation and Cell signaling	no subcategory	cAMP signaling in bacteria	cAMP-binding proteins - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases
Regulons	Atomic Regulons	ar-431-EC Molybdopterin-guanine dinucleotide biosynthesis	Molybdopterin-guanine dinucleotide biosynthesis protein MobB
Respiration	Electron accepting reactions	Anaerobic respiratory reductases	Arsenate reductase (EC 1.20.4.1)
Respiration	Electron accepting reactions	Anaerobic respiratory reductases	Vanillate O-demethylase oxidoreductase (EC 1.14.13.-)
Respiration	Electron accepting reactions	Terminal cytochrome O ubiquinol oxidase	Cytochrome O ubiquinol oxidase subunit I (EC 1.10.3.-)
Respiration	Electron accepting reactions	Terminal cytochrome O ubiquinol oxidase	Cytochrome O ubiquinol oxidase subunit II (EC 1.10.3.-)
Respiration	Electron accepting reactions	Terminal cytochrome O ubiquinol oxidase	Cytochrome O ubiquinol oxidase subunit III (EC 1.10.3.-)
Respiration	Electron accepting reactions	Terminal cytochrome O ubiquinol oxidase	Cytochrome O ubiquinol oxidase subunit IV (EC 1.10.3.-)
Respiration	no subcategory	Biogenesis of c-type	Periplasmic thiol: disulfide interchange

		cytochromes	protein DsbA
Stress Response	Detoxification	Uptake of selenate and selenite	Various polyols ABC transporter, ATP-binding component
Stress Response	Osmotic stress	Choline and Betaine Uptake and Betaine Biosynthesis	HTH-type transcriptional regulator BetI
Stress Response	Osmotic stress	Choline and Betaine Uptake and Betaine Biosynthesis	Sarcosine oxidase beta subunit (EC 1.5.3.1)
Stress Response	Oxidative stress	Glutaredoxins	Glutaredoxin 3
Stress Response	Oxidative stress	NADPH:quinone oxidoreductase 2	NADPH:quinone oxidoreductase 2
Stress Response	Oxidative stress	Oxidative stress	Ferroxidase (EC 1.16.3.1)
Stress Response	Oxidative stress	Oxidative stress	Iron-binding ferritin-like antioxidant protein
Stress Response	Oxidative stress	Oxidative stress	Non-specific DNA-binding protein *Dps*
Stress Response	Oxidative stress	Oxidative stress	Peroxidase (EC 1.11.1.7)
Stress Response	Oxidative stress	Oxidative stress	transcriptional regulator, Crp/Fnr family
Stress Response	no subcategory	Bacterial hemoglobins	Flavohemoprotein (Hemoglobin-like protein) (Flavohemoglobin) (Nitric oxide dioxygenase) (EC 1.14.12.17)
Sulfur Metabolism	no subcategory	Galactosylceramide and Sulfatide metabolism	Beta-galactosidase (EC 2.2.1.23)
Virulence, Disease and Defense	Resistance to antibiotics and toxic compounds	Arsenic resistance	Arsenic resistance protein ArsH
Virulence, Disease and Defense	Resistance to antibiotics and toxic compounds	Beta-lactamase	Beta-lactamase class C and other penicillin binding proteins
Virulence, Disease and Defense	Resistance to antibiotics and toxic compounds	Cobalt-zinc-cadmium resistance	Cation efflux system protein CusA
Virulence, Disease and Defense	Resistance to antibiotics and toxic compounds	Cobalt-zinc-cadmium resistance	Cobalt-zinc-cadmium resistance protein *CzcA*

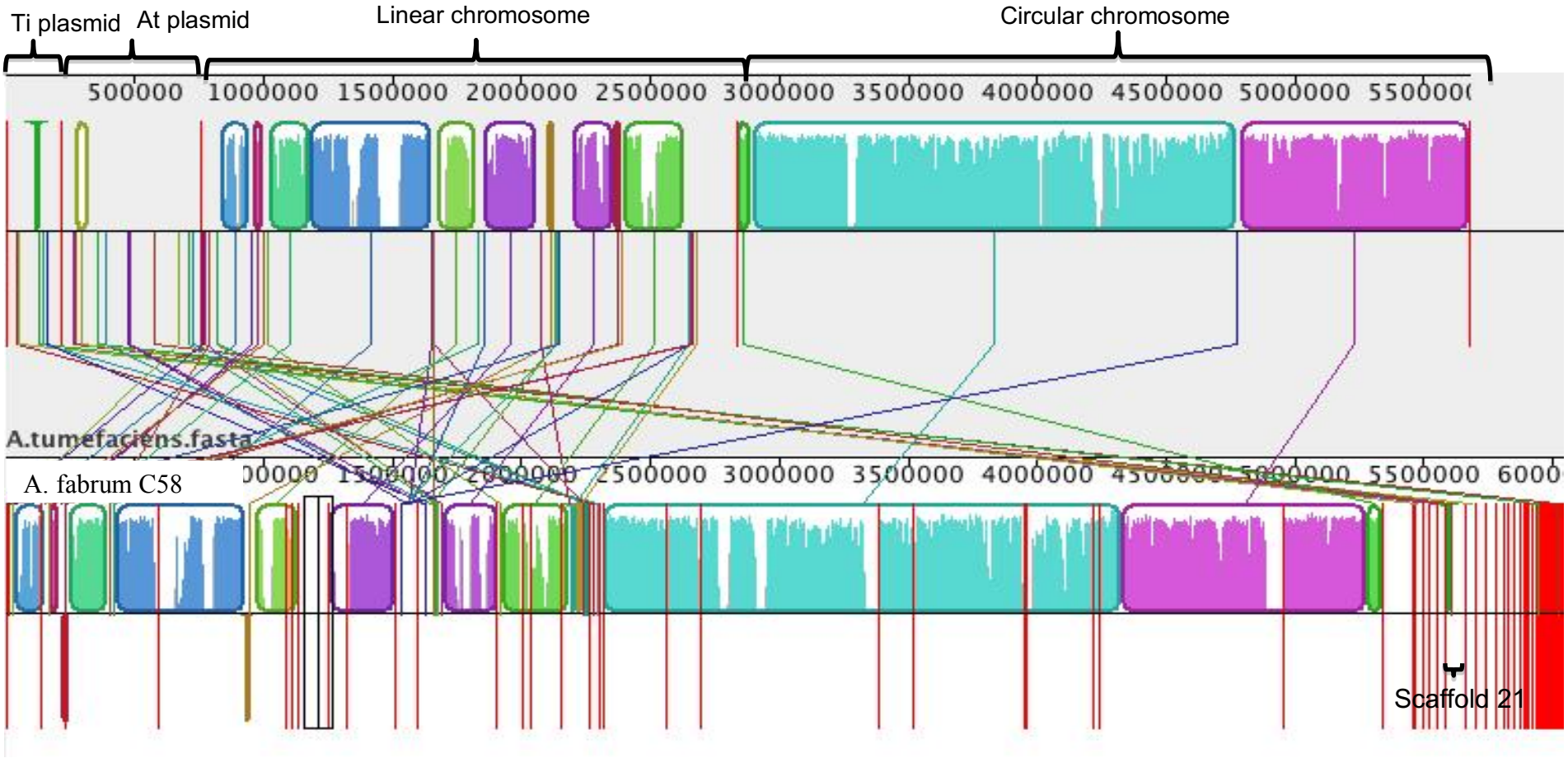


Figure 2.23 Mauve alignment between *A. fabrum* C58 (top) and *Agrobacterium* sp. SUL3 (bottom). Vertical red lines indicate scaffold boundaries in *A. tumefaciens* SUL3 and boundaries between the Ti plasmid, At plasmid, linear chromosome and circular chromosome in *A. fabrum* C58 (labelled). Areas of homology between the two sequences are represented with coloured blocks There is a small region of similarity between the Ti plasmid of *A. fabrum* C58 and a region of *Agrobacterium* sp. SUL3 (scaffold 21). Large regions of similarity arepresent between the circular chromosome of *A. fabrum* C58 and *Agrobacterium* sp. SUL3 and smaller regions of similarity are present between the linear chromosome and *Agrobacterium* sp. SUL3.

Table 2.11 Top ten BLAST hits for *Agrobacterium* sp. SUL3 scaffold 21. Ordered by maximum score according to NCBI BLAST

Description	Query coverage (%)	Identity (%)
Rhizobium sp. IRBG74 plasmid IRBL74	32	83
Rhizobium leguminosarum bv. *trifolii* WSM2304 plasmid pRLG203	32	79
Ochrobactrum anthropi ATCC 49188 plasmid pOANT02	33	88
Rhizobium etli bv. *mimosae* str. *Mim1* plasmid pRetMIM1e	49	77
Rhizobium leguminosarum bv. *trifolii* WSM1689 plasmid	33	79
Rhizobium leguminosarum bv. *viciae* strain VF39 plasmid pRleVF39b	30	80
Agrobacterium rhizogenes plasmid pRi1724	19	76
Agrobacterium tumefaciens F64/95 plasmid pAoF64/95	19	76
Agrobacterium rhizogenes plasmid pRi2659	19	78
Polymorphum gilvum SL003B-26A1 plasmid pSL003B	17	77

relationship is largely determined by the presence of plasmids, with beneficial symbionts having *nod* and *nif* genes on symbiotic plasmids whilst pathogenic strains have *vir* genes on Ti plasmids (Valazquez *et al*., 2005). Using both RAST and BLAST *nod, nif* and *vir genes* were all searched for in SUL3. No evidence was found for any of these genes; however, RAST identified a number of conjugative transfer proteins present in SUL3, on scaffolds 21, 23 and 26. Conjugative transfer proteins facilitate the horizontal gene transfer of mobile elements and are normally found on the Ti plasmid in *A. tumefaciens.* By zooming in on the Mauve alignment it can be seen that scaffolds 21, 23 and 26 all correspond to a small region of similarity between *Agrobacterium* sp. SUL3 and the Ti plasmid of *A. fabrum* C58 (figure 2.24). This indicates that *Agrobacterium* sp. SUL3 has partial plasmid elements present; however, essential *vir* genes are absent meaning it is unlikely to be pathogenic.

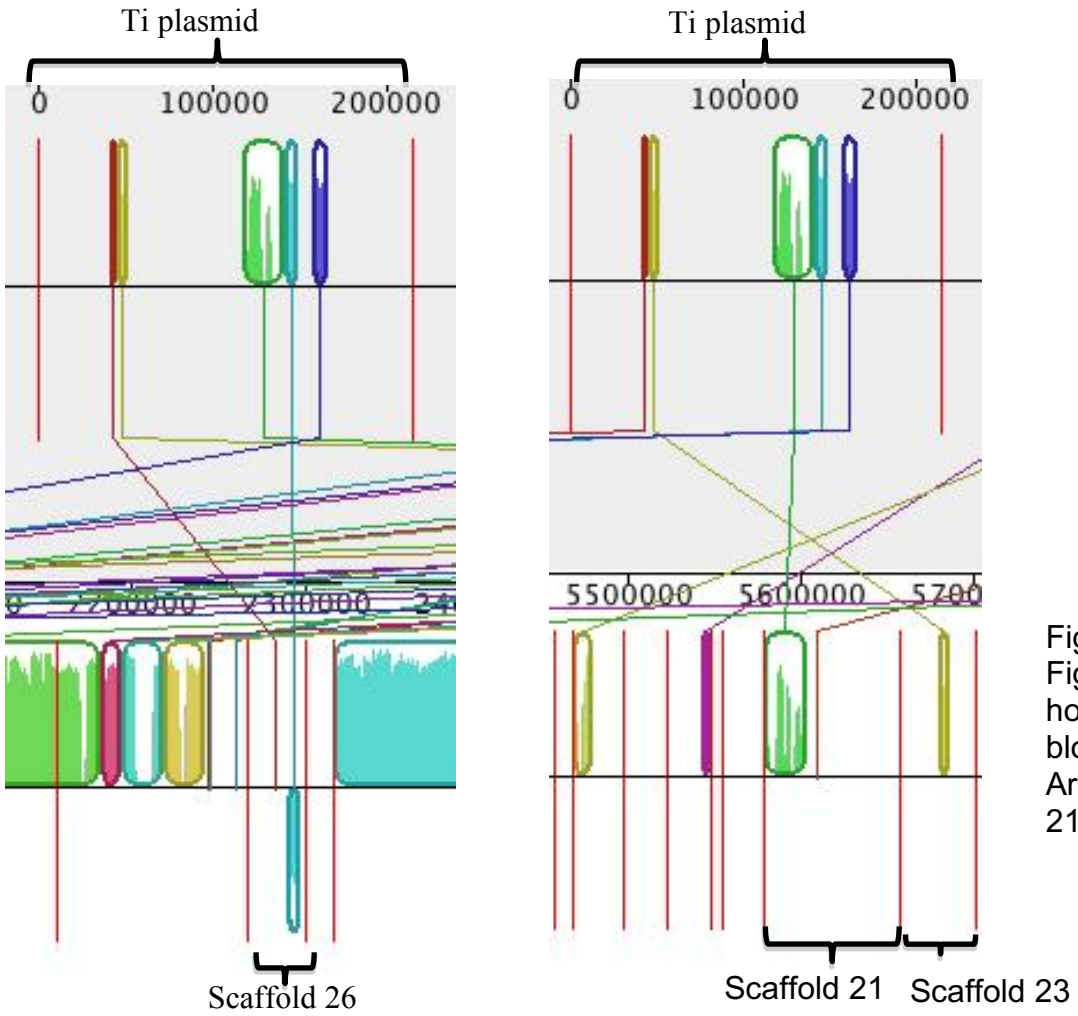


Figure 2.24 Zoomed in region of the Mauve alignment shown in Figure 2.22, focussing on the Ti plasmid of *A. fabrum* C58. Areas of homology between the two sequences are represented with coloured blocks and scaffold boundaries are represented by red vertical lines. Areas of similarity are present between the Ti plasmid and scaffolds 21, 23 and 26 in *Agrobacterium* sp. SUL3

2.7 Bacterial strain GCS4 is a member of the genus *Microbacterium*

2.7.1 Phylogenetic analysis indicates bacterial strain GCS4 is a species of *Microbacterium*

As with the previously discussed bacterial strains, the 16S rRNA gene sequence was extracted from the genome of bacterial strain GCS4 using RNAmmer and used as the query in a BLAST search against the NCBI 16S ribosomal DNA sequences (bacteria and archea) database. The BLAST search returned results exclusively from the family Microbacteriaceae, the majority of which belonged to the genus *Microbacterium.* Phylogenetic analysis was carried out using the 16S rRNA gene sequence from bacterial strain GCS4 along with sequences from the BLAST search which showed over 95% identity. The phylogenetic analysis placed bacterial strain GCS4 in a clade with *Microbacterium hydrocarbonoxydans* (Figure 2.25). In order to try and further determine which species of *Microbacterium* bacterial strain GCS4 is likely to be, a number of conserved genes (*gyrB, ppk, recA* and *rpoB*) were used for further phylogenetic analysis. These genes were previously used in a study by Richert *et al.* (2007) to construct phylogenies of 27 type strains of *Microbacterium*. The phylogenetic tree constructed using the four conserved genes shows bacterial strain GCS4 in a clade with *M. foliorum* and *M. phyllosphaerea* (Figure 2.26), which were more distant in the 16S rRNA phylogenetic tree. A number of *Microbacterium* species which appeared phylogenetically close according to the 16S rRNA tree are not included in the second tree as sequences were not available for them. The phylogenetic analysis allows for bacterial strain GCS4 to be classified to the genus level, and it will be referred to as *Microbacterium* sp. GCS4 in all subsequent sections.

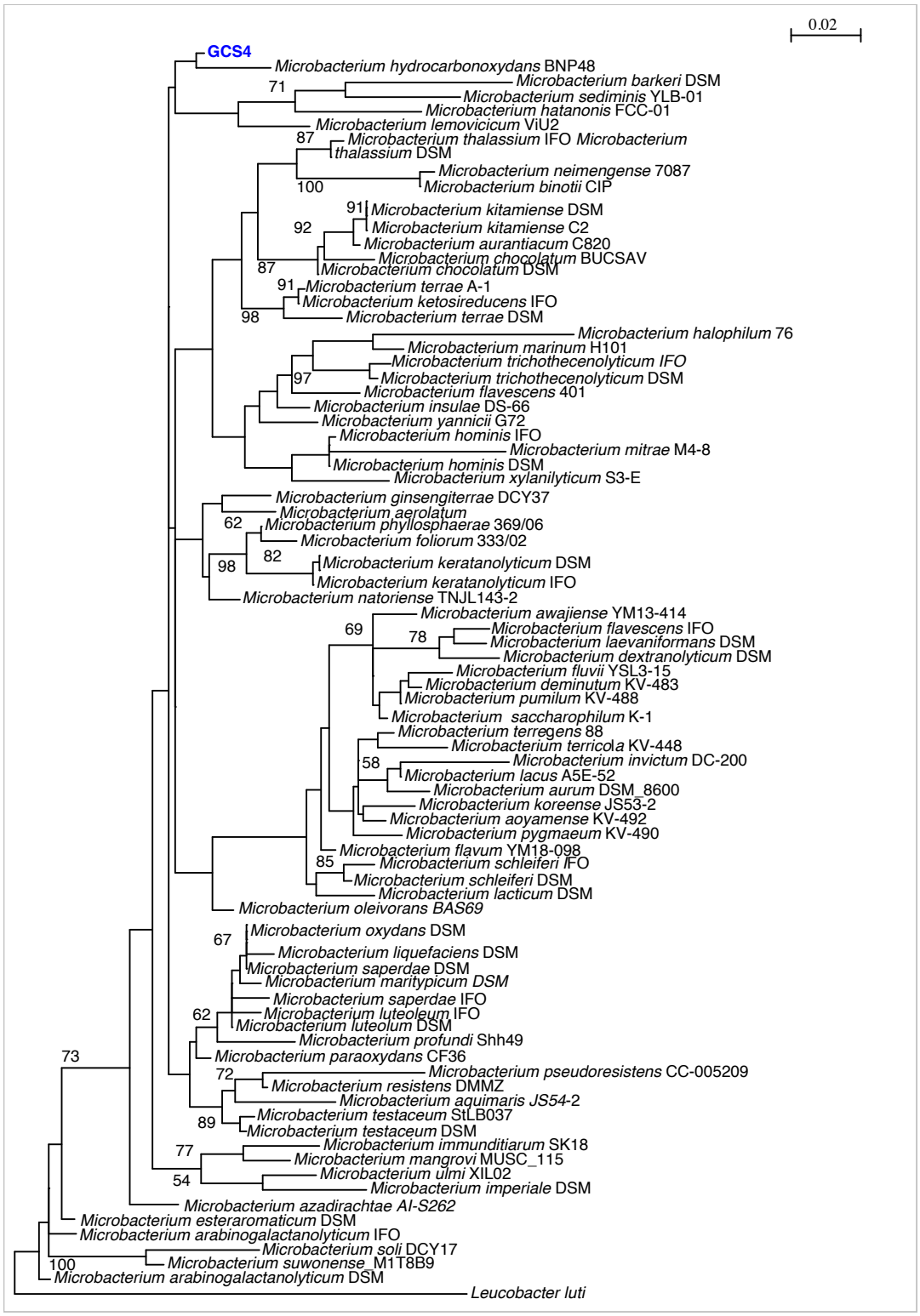


Figure 2.25 Phylogram constructed from maximum likelihood analysis (PhyML) of 16S rRNA gene sequence data for bacteria, identified through BLAST as having > 95% identity to the16S rRNA gene sequence of bacterial strain GCS4. Bootstrapping was set at 100 and values are shown for each node with > 50% bootstrap support. The tree is rooted with *Leucobacter luti* as

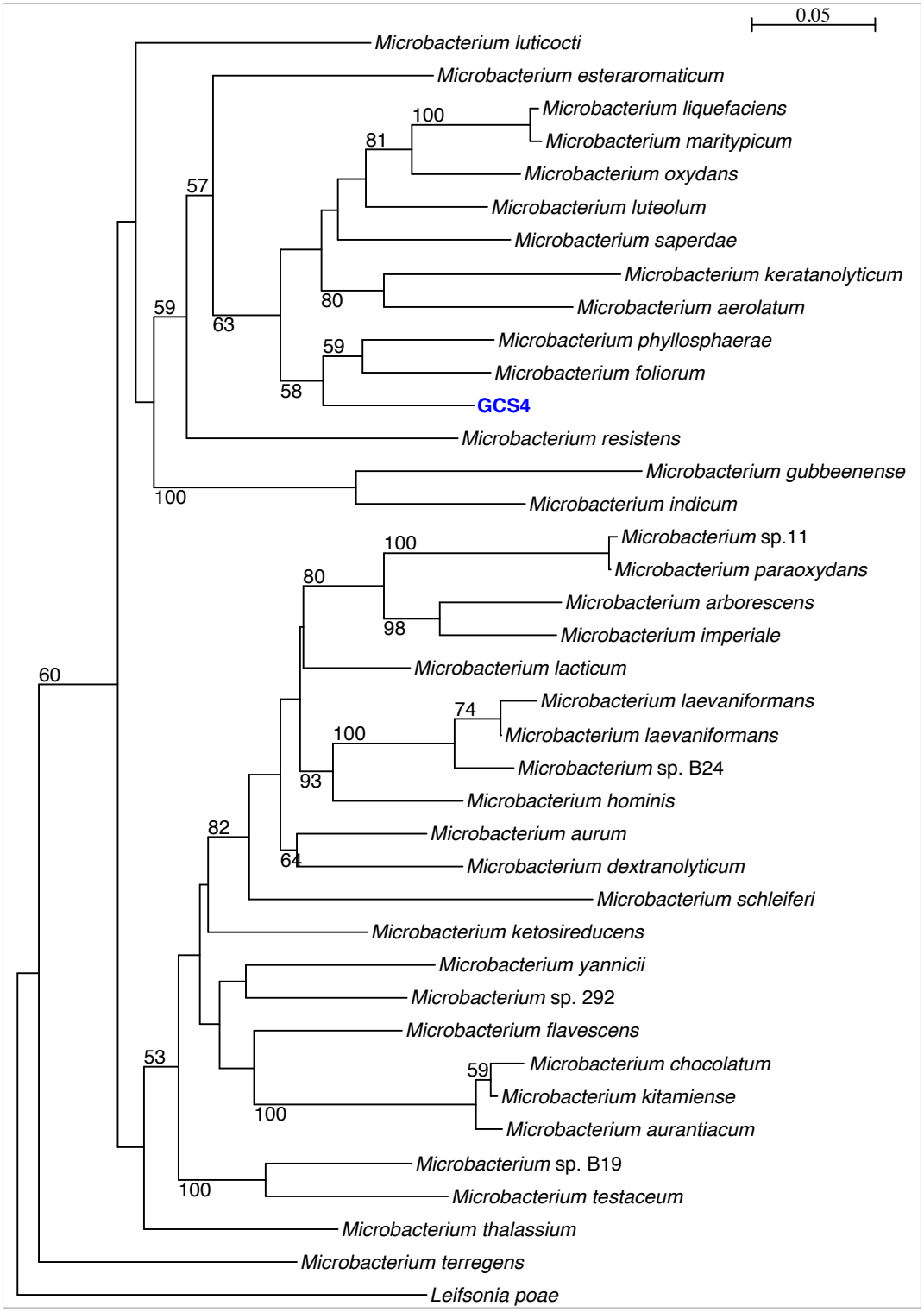


Figure 2.26 Phylogram constructed from maximum likelihood analysis (PhyML) of *recA*, *rpoB, tyrb* and *ppk* sequence data for bacterial strain GCS4 and bacteria from the genus *Microbacterium*. Bootstrapping was set at 100 and values are shown for each node with > 50% bootstrap support. The tree is rooted with *Leifsonia* poae as the outgroup.

2.7.2 Whole genome comparisons indicate differences between GCS4 and other *Microbacterium* species

Due to uncertainties regarding the closest relative of GCS4, all *Microbacterium* genomes available on the NCBI database were used in a BRIG whole genome comparison (Figure 2.27). The BRIG diagram indicates that there are a large amount of differences between GCS4 and the species of *Microbacterium* which have genomes available in the NCBI database. The 16S rRNA phylogenetic analysis of *Microbacterium* sp. GCS4 places it in a clade with a strain of *Microbacterium hydrocarbonoxydans* previously isolated from oil contaminated soil in Germany where it was found to be crude-oil degrading (Schippers *et al*., 2005). In order to assess whether *Microbacterium* sp. GCS4 also potentially has these oil degrading properties we searched for *alkB*, a marker frequently used to asses the alkane degrading potential of bacteria. The *alkB* gene encodes for alkane 1-monooxygenase, a key enzyme in the initial oxidation of alkanes (Sheng *et al*., 2009). A BLAST alignment using the tblastn algorithm (protein-sequence query against a translated nucleotide database) using the amino acid sequence for the alkane 1-monooxygenase found in *M. hydrocarbonoxydans* as the query against the genome of *Microbacterium* sp. GCS4 returned an alignment with 100 % query coverage, an E-value of 0 and a 90 % identity to the same region on scaffold number four. This would indicate that there is an *alkB*-like gene present in *Microbacterium* sp. GCS4. The whole genome sequences of *M. hydrocarbonoxydans* and *Microbacterium* have an ANI of 85.5%, indicating two different species.

It has not been possible to identify GCS4 to species level using whole genomes and other gene sequences currently available. This would indicate that GCS4 is a species of *Microbacterium* that has not previously been extensively studied. Genes of interest (which may be indicate interactions between the bacteria and *B. braunii*) including those involved in vitamin synthesis, secretion systems and nitrogen fixation were looked for within the genome sequence of *Microbacterium sp.* GCS4; these are discussed in more detail in sections 2.8 – 2.10.

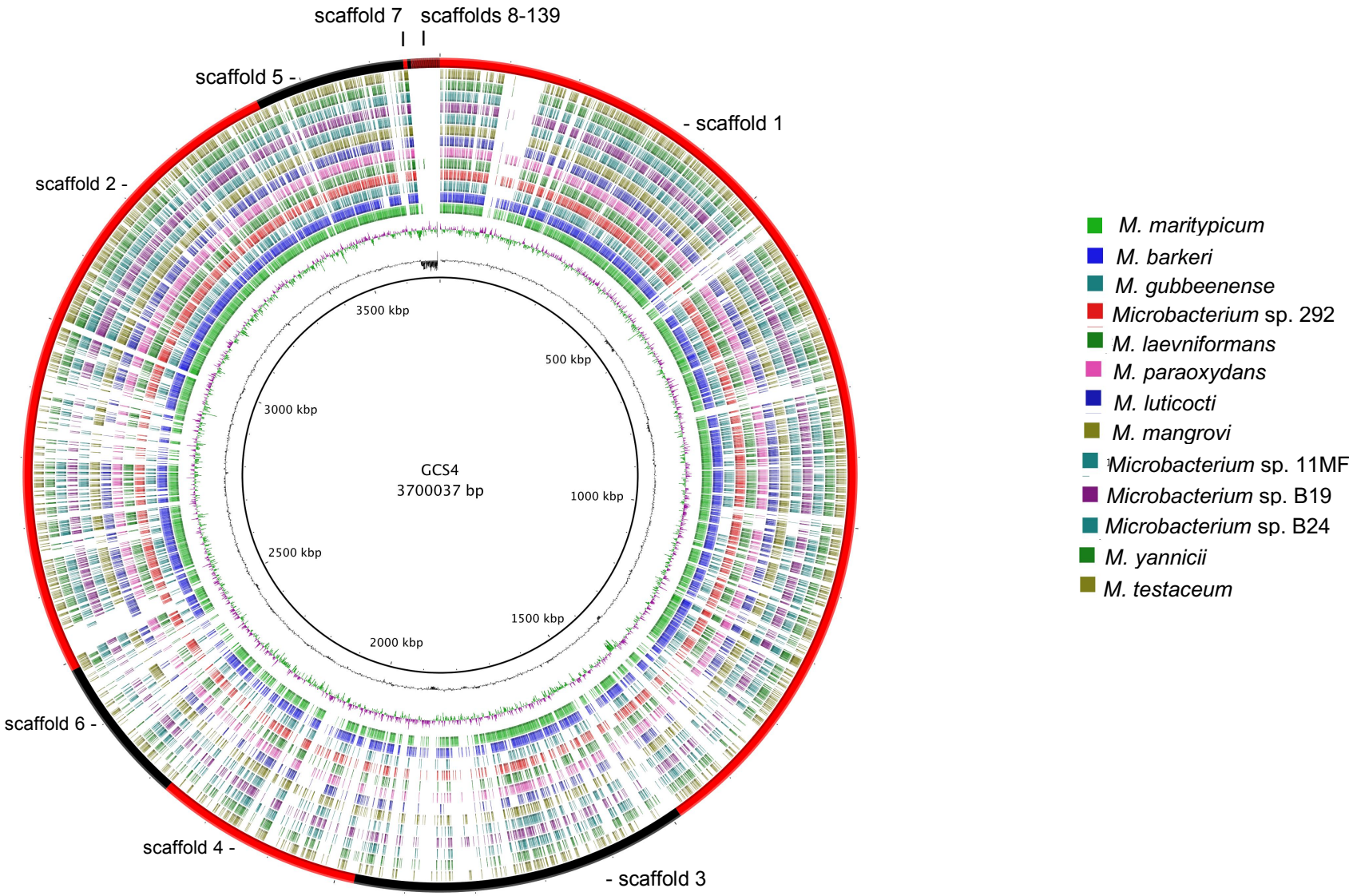


Figure 2.27 Whole genome comparisons between *Microbacterium* GCS4 and thirteen other *Microbacterium* species, created using the BLAST Ring Image Generator (BRIG.) The alternating red and black circle indicates *Microbacterium* GCS4 scaffolds, whilst inner coloured circles give a graphical representation of areas of homology between the reference sequence (*Microbacterium* GCS4) and the query sequences (the *Microbacterium* species).

2.8 B vitamin synthesis pathways are present in all the *B. braunii* consortium bacterial strains

Many species of algae are auxotrophic for vitamin B1 (thiamine), vitamin B7 (biotin) and vitamin B12 (cobalamin) (Croft *et al*., 2006). Auxotrophic algae must obtain all or some of these vitamins from the environment; however the majority of algal habitats contain insufficient levels of these B vitamins to support their growth (Karl, 2002). Auxotrophic algae must therefore be acquiring B vitamins from another source, and studies have shown this is most probably through a symbiotic relationship with bacteria which are synthesising these vitamins (Croft *et al*., 2005). Although studies have been carried out to determine which species of algae are auxotrophic for B vitamins (Tang *et al*., 2010) there is no information available regarding vitamin requirements of *B. braunii*. Algal species with a requirement for these vitamins are found across a number of unrelated phyla and it is therefore hard to make predictions regarding which species may be auxotrophs. We can therefore only hypothesise that *B. braunii* requires some or all of these vitamins which may be synthesised by members of the *B. braunii* bacterial consortium.

Whole genome analysis and annotation using RAST was used to look for B vitamin synthesis pathways (Table 2.12). This shows that thiamine is likely to be synthesised by all the *B. braunii* consortium bacterial strains, cobalamin is likely to be synthesised by *A. piechaudii* GCS2 and *Shinella* sp. GWS1/SUS2, whilst biotin is likely to be synthesised by *A. piechaudii* GCS2 and *Agrobacterium* sp. SUL3 All five bacteria may therefore be benefiting *B. braunii* through various B vitamin synthesis, with *A. piechaudii* GCS2 potentially the most beneficial as it has pathways present for all three of the B vitamins considered.

Table 2.12 B vitamin synthesis pathways in the *B. braunii* consortium bacterial strains

Bacterial strain	**Biotin (Vitamin B7) biosynthesis**	**Thiamine (Vitamin B1) biosynthesis**	**Cobalamin (Vitamin B12) biosynthesis**
Achromobacter piechaudii GCS2	+	+	+
Microbacterium sp. GCS4	-	+	-
Shinella sp. GWS1 and SUS2	-	+	+
Agrobacterium sp. SUL3	+	+	-

Table 2.13 Secretion systems present in the *B. braunii* consortium bacterial strains.

Bacterial strain	**Type I SS**	**Type II SS ***	**Type III SS**	**Type IV SS**	**Type V SS**	**Type VI SS**
Achromobacter piechaudii GCS2	-	+	-	-	-	+
Microbacterium sp. GCS4	-	+	-	-	-	-
Shinella sp. GWS1/SUS2	-	+	-	+	-	+
Agrobacterium sp. SUL3	-	+	-	+	-	+

*This refers to a subsystem of the type II secretion system: the Tad (tight adherence) macromolecular transport system

2.9 Secretion systems indicate possible interactions between the bacteria and the alga

Protein secretion systems are a key factor in how a bacterial species is able to interact with its environment, playing a crucial role in pathogenic, commensal or mutualistic relationships between bacteria and eukaryotic host organisms (Tseng *et al*., 2009). There are six known secretion systems which occur in Gram negative bacteria (T1SS-T6SS). (Gerlach and Hensel, 2007). Using RAST, secretion systems were looked for in each of the five bacterial strains isolated from *B. braunii*; Table 2.13 shows a summary of which secretion systems are present in each bacterial strain.

A sub-type of the type II secretion system known as the Tad (tight adherence) macromolecular transport system is encoded in all of our bacterial genomes. A number of *tad* genes are present on a genomic island known as the widespread colonisation island. The *tad* genes are responsible for the assembly of adhesive Flp (fimbrial low-molecular-weight protein) pili which form long, filamentous fibrils which mediate biofilm formation, pathogenesis and colonisation of solid surfaces in a range of bacteria (Tomich *et al.,* 2007). The presence of this secretion system in all of our samples indicates they have the ability to form biofilms and could be colonising the algal surface rather than living freely in the water column. The presence of bacteria in a biofilm on *B. braunii* has been previously observed by Rivas *et al.* (2010).

The type IV secretion system is present in *Agrobacterium* sp. SUL3 and *Shinella* sp SUS2/GWS1. Type IV secretion systems are diverse, and may carry out a range of functions including the transfer of virulence factors by pathogenic bacteria and the mediation of horizontal gene transfer (Wallden *et al*., 2010). The type IV secretion system present in *Agrobacterium* sp. SUL3 and *Shinella* sp. GWS1/SUS2 is related to bacterial conjugation systems. Bacterial

conjugation systems enable the transfer of mobile elements between bacteria, enhancing the spread of resistance, virulence and social traits amongst prokaryotes (Guglielmini *et al*., 2013). Consortium members with this secretion system present may therefore be benefiting from the acquisition of antibiotic resistance genes or virulence related genes (Vogan and Higgs, 2011). This is highly likely to be due to the *B. braunii* culture from which the bacteria were isolated having been a lab culture for an extended period of time and therefore likely to have been exposed to numerous antibiotics. *Shinella* sp. GWS1/SUS2 and *Agrobacterium* sp. SUL3 have a number of genes encoding beta-lactamases - enzymes that confer resistance to a wide range of beta-lactam antibiotics (Colodner *et al*., 2014) - as well as genes conferring resistance to fluoroquinolones and streptothricin. Additionally, *Agrobacterium* sp. SUL3 and *Shinella* sp. GWS1/SUS2 encode multidrug resistance efflux pumps which may have the dual role of conferring antibiotic resistance as well as resistance to natural substances produced by bacterial hosts (Piddock, 2006) – in this case by *B. braunii.*

The recently discovered type VI secretion system, present in all the bacterial genomes except *Microbacterium* sp. GCS4, has been linked to virulence in a range of bacterial pathogens which target eukaryotic hosts, as well as to interactions between other members of a bacterial population. The type VI secretion system allows interaction between bacteria through the injection of effector proteins into target cells. Type VI secretion systems can be placed into two broad categories: those that target eukaryotic cells and those that target bacterial cells (Russell *et al*., 2014). Within bacterial communities it has been suggested that this secretion system is used as weapon against competitors by injection of antibacterial toxins into other bacterial cells (Coulthurst, 2013). The presence of a system which may prove beneficial within a large community is likely to give bacterial strains GCS2, GWS1/SUS2 and SUL3 an advantage over competitor bacteria which may not possess this system.

Type I, III and V secretion systems are not present in any of our samples. Significantly, the type III secretion system is associated with pathogenesis and virulence in Gram-negative bacteria where it allows for effector proteins to be injected directly into bacterial hosts (Coburn *et al*., 2007). The absence of this

secretion system in all the bacterial strains suggests they are not pathogenic to *B. braunii.*

2.10 Nitrogen fixation genes and alkane utilisation pathways searched for in all the *B. braunii* bacterial consortium genomes

It has been hypothesised that nitrogen fixation is one way in which bacteria may be beneficially interacting with algae (Chirac *et al*., 1985). Nitrogen fixation (*nif*) genes were looked for in all the *B. braunii* bacterial consortium genomes using both RAST and BLAST. No evidence was found for any nitrogen fixation genes in any of the genomes.

As well as showing that nitrogen fixation is not one of the ways these bacterial strains are interacting with *B. braunii*, the lack of *nif* genes also provides further evidence to support a number of the taxonomic classifications that have been previously made in this study. *Shinella* sp. GWS1/SUS2 appears to be a strain of *Shinella*, the majority of which do not have nitrogen fixing properties. (As discussed in section 2.5).

Botryococcus braunii produces high levels of hydrocarbons, the majority of which are alkenes. It was therefore of interest to determine if the bacteria we have isolated from *B. braunii* may utilise these hydrocarbons. The *alkB* gene has been identified as being essential to hydrocarbon biodegradation (Hassanshahian *et al*, 2013). Therefore, *alkB* genes, along with *alkM* genes which encode for alkane hydroxylase, were searched for within all five genomes using BLAST. As previously discussed, evidence for the *alkB* gene was found in *Microbacterium* sp. GCS4; however, no sequences encoding *alkB* were found in any of the other genomes. No evidence of the *alkM* gene was found in any of the bacterial strains. However, genome annotations analysis indicated the presence of an aromatic hydrocarbon utilisation transcriptional regulator *CatR* in

A. piechaudii GCS2, *Microbacterium* sp. GCS4 and *Shinella* sp. GWS1/SUS2. Aromatic hydrocarbons are also produced by *B. braunii* (Banerjee *et al*., 2001) and it is therefore possible the bacteria may be utilising these.

2.11 Summary

Next generation sequencing technology and a variety of bioinformatics tools have been used to successfully sequence the genomes of five bacterial strains isolated from consortium with the eukaryotic alga *Botryococcus braunii* and to identify these bacteria. Phylogenetic analysis using both 16S rRNA gene sequences and housekeeping genes has enabled one bacterial strain to be assigned to the species level and four to genus level. Whole genome analysis of metabolic pathways predicts secretion systems and vitamin synthesis pathways indicating a range of interactions are occurring between the bacteria and *B. braunii.*

In addition to the work detailed so far, we have also extracted and sequenced the 16S rRNA gene sequences from the entire microbial community found in consortium with *B. braunii* and this dataset will now be analysed to further determine the complexity of the community.

Chapter three: Analysing the microbial consortium of *Botryococcus braunii* using 16S rRNA gene sequencing

3.1 Introduction

Chapter three detailed the genomic analysis of five bacterial strains isolated from *Botryococcus braunii.* Whole-genome analysis allows for detailed taxonomic classification and metabolic reconstructions. However, it cannot answer questions regarding how representative these bacterial strains are of the bacterial population living alongside *B. braunii*, nor can it be determined if these bacterial strains are usually found living in close or loose association with their algal host. These questions are addressed in this chapter by analysing the community structure of the bacterial population found living in close association with *B. braunii* and the bacterial population living in looser association, within the water column. This is achieved through the extraction, amplification and sequencing of 16S rRNA genes from the bacterial community.

3.1.1 The use of 16S rRNA sequencing for microbial community analysis

The most long-established molecular method of identifying members of a microbial ecosystem is the sequencing of 16S rRNA. Carl Woese and George Fox pioneered the use of 16S rRNA in taxonomic classification in 1977. Woese and Fox recognised that ribosomal RNA was central to cellular function with a nucleotide sequence that changes very slowly over time, enabling the relatedness of distant species to be determined (Woese and Fox, 1977). Using 16S rRNA gene sequence analysis Woese and Fox revolutionised the biological world by classifying archaea as phylogenetically distinct from bacteria and thereby defining that there were in fact three domains in the tree of life and not two as it had previously been believed. Following on from this, Norman Pace's seminal work in 1985 pioneered the use of 16S rRNA gene sequencing as a tool for directly identifying microbial populations in environmental samples and by the early 1990s studies analysing bulk 16S rRNA gene sequences from environmental samples were becoming more widespread (Lane *et al*., 1985; Giovannoni *et al*., 1990; Schmidt *et al*., 1991). The use of the 16S rRNA gene for taxonomic classification remains popular today and there are now a large number of tools available for analysis as well as extensive databases for sequence comparisons (Clarridge, 2004; Janda & Abbott, 2007).

The 16S rRNA gene codes for the RNA component of the 30S subunit of the prokaryotic ribosome and is present in all prokaryotes. The gene contains nine hypervariable regions interspaced between conserved regions (Van de Peer *et al*., 1996) (Figure 3.1). As the conserved regions flank the variable regions it is possible to carry out PCR amplification of targeted variable regions (Chakravorty *et al*. 2007). Studies have demonstrated differences in the usefulness of different variable regions for taxonomic classification depending on the bacteria or environment of interest, though it should be noted that the majority of these studies have focused on bacteria of interest to clinical research as opposed to environmental samples (Chakravorty *et al*. 2007; Kumar et al., 2011). However, no single variable region can be used to differentiate between all bacteria and a certain level of bias is therefore introduced into bacterial community profiles depending on which variable region is selected (Kumar *et al*., 2011).

Despite its widespread use, there are a number of limitations when using 16S rRNA gene sequences for taxonomic classification. The resolving power of 16S rRNA gene sequence analysis at and below the species level is poor (Martens *et al*., 2008); the high levels of similarity between 16S rRNA sequences from different species mean a minimal 16S rRNA gene sequence similarity value for determining species cannot be set (Jaspers & Overmann, 2004) When comparing bioinformatics pipelines for the analysis of 16S rRNA gene sequence data (see 3.1.2, below) Plummer *et al*. (2015) also highlighted the limitations of 16S rRNA for genus and species level identification, drawing attention to the need for caution when interpreting output from analysis tools. The quality of 16S rRNA sequence databases must also be considered; complete, unambiguous and correctly labelled nucleotide sequences within a database are key elements in taxonomic classification (Janda & Abbott, 2007). Historically many depositions in 16S rRNA sequence databases were of poor quality, with large numbers of sequencing errors and chimeras (Ashleford *et al*., 2005), additionally, a large number of 16S rRNA gene sequences in the GenBank database have been labelled “environmental samples” or “unclassified” (DeSantis *et al*., 2006). However, a number of databases have addressed these issues: the Greengenes 16S rRNA database provides features to mitigate

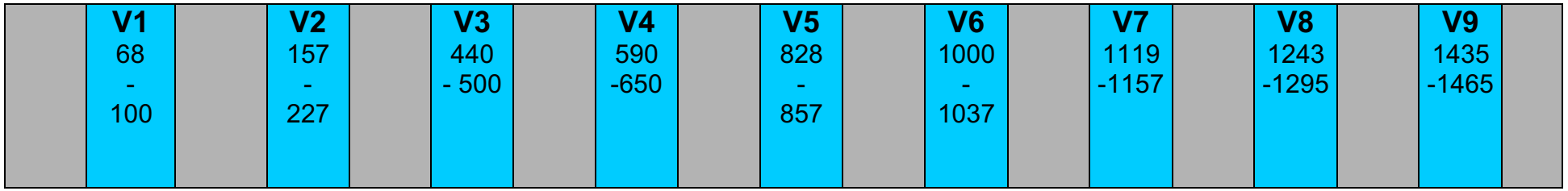


Figure 3.1 16S variable and conserved regions (in blue and grey respectively) and their approximate base pair positions in *E. coli* (According to Thomas *et al*., 2011)

Incompatible taxonomic nomenclature amongst database curators as well as chimera screening (DeSantis *et al*., 2006), the SILVA rRNA database also carries out an ever-increasing number of quality control measures on sequences deposited there, since its release (Pruesse *et al*., 2007; Quast *et al.,*2012) and the NCBI 16S RefSeq collection carries out sequence validation steps including low quality sequence removal, vector screening and chimera checking (NCBI, 2017).

3.1.2 Tools for analysing microbial communities

Studies of microbial communities using metagenomics or 16S rRNA gene sequence data have become increasingly widespread, with a diverse range of environments targeted. As a consequence of this, a number of bioinformatics tools have been developed to analyse and taxonomically classify the large sequencing datasets generated from these studies (Peabody *et al*., 2015). These tools vary in both function and the amount of interaction required from the user, with some utilising graphical user interfaces and others requiring extensive use of the command line (Lindgreen *et al*., 2016). These bioinformatics tools can be further divided into two broad categories: those which analyse an entire metagenomics data set, and abundance estimation programs which analyse a smaller representative set of sequences, such as 16S rRNA and other phylogenetic marker genes (Wood and Salzberg, 2014).

Accuracy of the taxonomic classification carried out by bioinformatic classification tools is essential in order to draw conclusions regarding the composition of a microbial community, however, studies into the accuracy of these tools is limited. Nilikanta *et al.* (2014) published a review of seven software packages: Mothur, QIIME, W.A.T.E.R.S, RDPipeline, VAMPS, Genboree and SnoWMan these tools met their inclusion criteria of being free, publicly available, offering analysis functions from platform sequencing to results presentation, and have documentation and data security. The study did not test the output of any of the software packages but instead reviewed their installation, documentation, functions and features, coming to the conclusion that all packages were likely to perform well, but that the packages Mothur and QIIME stood out as “outstanding” due to their suite of functions and features as well as their good documentation. Plummer *et al*. (2015) compared the performance of three bioinformatics tools - QIIME, mother and MG-RAST for

the analysis of gut microbiota 16S rRNA gene sequences. The three methods were compared with regards to taxonomic classification, diversity analysis and usability; the study concluded that QIIME and mothur had superior statistical capabilities and user freedom over MG-RAST, with QIIME being more user friendly than Mothur. However, the study also acknowledged the limitations of using 16S rRNA gene sequences for taxonomic classification, highlighting the problems that arise when different species have highly similar sequences and suggesting that even genus level identification can be unreliable.

A study by Lindgreen *et al*. (2016) used a synthetic metagenomic community to look at the accuracy and speed of fourteen metagenome analysis tools and judged their performance based on a number of factors. The study used shotgun metagenomic sequence data. However, the tools discussed can also be used for the analysis of 16S rRNA gene sequence data. The study found that all of the tools generated relative abundance figures that differed from the actual abundances present in the community although they were variable in terms of the levels of divergence. However, when assessing false positives and shuffled genomes (Shuffled reads were obtained by shuffling a set of 110 genomes, using the HMMER shuffle program, to mimic a pool of unknown reads which should not be mapped to any taxa) a number of tools clearly outperformed others. Table 3.1 shows a summary of these results and Table 3.2 gives further information on their use.

3.1.3 Aims

This chapter will address questions regarding whether the bacterial strains detailed within chapter two are in close or loose association with *B. braunii* and if they are abundant within the community, as well as providing greater insight into the general community structure of bacterial populations found in close and loose association with *B. braunii*. Additionally, this chapter will address differences in the results of taxonomic distribution analysis depending on which variable region of the 16S rRNA gene has been used in order to determine the best methodologies to use in future studies.

Table 3.1 Phylum level performance metrics for fourteen methods of analysing metagenomics sequence data. Fraction: average fraction of all reads that the tool mapped. Shuffled: average number of shuffled reads mapped. False positives: fraction of mapped reads assigned to non-existing phyla. Run time: CPU time in minutes per metagenome (where applicable). Taken from Lindgreen *et al*. (2015)

Analysis tool	**Fraction**	**Shuffled**	**False positives**	**Run time**
CLARK	73.32 %	340,607	0.02 %	211.50
EBI	0.08 %	0	41.74 %	~12 days
Genometa	39.91 %	0	0.83 %	401
GOTTCHA	43.10 %	NA	0.00 %	229.49
Kraken	71.98 %	19	0.00 %	60.95
LMAT	56.61 %	1,486,699	0.63 %	981.21
MEGAN	42.21 %	NA	0.49 %	2489.65
MetaPhlAn	5.09 %	0	0.75 %	108.51
MetaPhyler	0.45 %	649	0.05 %	26586.15
MG-RAST	56.17 %	3	0.27 %	16881.8
mOTU	0.16 %	NA	0.10 %	45.8
One Codex	73.68 %	23	0.00 %	27.77
QIIME	58.23 %	0	0.28 %	8.88
Taxator-tk	45.67 %	2	14.07 %	9147.92

Table 3.2 Metagenomic analysis tools used in the study by Lindgreen *et al* (2016) with additional information showing whether they use a graphical interface or command line and whether they analyse all sequences in a data set or only representative marker genes.

Analysis tool	Graphical Interface	Command line	Whole data set	Representative data set
CLARK		✓	✓	
EBI	✓*		✓	
Genometa	✓		✓	
GOTTCHA		✓	✓	
Kraken		✓	✓	
LMAT		✓	✓	
MEGAN	✓	✓	✓	
MetaPhlAn		✓		✓
MetaPhyler		✓		✓
MG-RAST	✓*		✓	
mOTU		✓		✓
One Codex	✓*		✓	
QIIME		✓		✓
Taxator-tk		✓	✓	

* Web based GUI

3.2 Materials and methods

3.2.1 Culturing of *Botryococcus braunii*

200 ml of *B. braunii* culture was sieved in a 20 µm sieve and rinsed with Milli-Q water. The flow-through was discarded and the remaining alga was removed from the sieve, placed into 50ml Falcon tubes and centrifuged at 4600 rpm, 20 ^{o}C for ten minutes. Supernatant containing *B. braunii* was added to 150 ml of fresh Chu 13 medium in a 1 litre conical flask and incubated in a shaking incubator with CO_2 at 5 %, shaking at 90 rpm, at 23 °C, photoperiod set to 18 hours light 6 hours dark for seven days. The culture was washed and placed in fresh medium (following the above protocol) again after 7 days and 14 days and DNA extracted after 21 days.

3.2.2 DNA extraction and sequencing from two fractions of *B. braunii*

DNA was extracted from two fractions of the *B. braunii* culture in order to assess which bacteria are present in the water column (Sample A) and which bacteria are present in close association with the alga (Sample B).

180 ml of cultured *B. braunii* was sieved in a 20 µm sieve. The flow-through was placed in 50 ml Falcon tubes, centrifuged at 4600 rpm for ten minutes, pellets were combined into one microfuge tube, centrifuged for five minutes at 1200 rpm and washed a further two times with MilliQ water to give sample A. The alga was eluted from the sieve using Milli-Q water into a 50 ml Falcon tube, sonicated in a water bath for ten minutes and centrifuged at 4600 rpm for ten minutes; the pellet was centrifuged and washed a further two times to give sample B.

Genomic DNA extraction was carried out using the Sigma-Aldrich Bacterial Genomic Miniprep kit, following the standard Gram-negative bacteria protocol.

Three variable regions of bacterial 16S rRNA were targeted for amplification, using primers designed for the V1-V2, V3-V4 and V5-V8 variable regions. PCR reaction mix was created using the NEBnext High-Fidelity PCR master mix. A second PCR phase was incorporated to add flowcell binding regions, an illumina adapter and a multiplexing barcode, creating DNA libraries which were combined

and sequenced using the Illumina MiSeq. 300bp paired end reads were obtained and demultiplexed by The Exeter Sequencing Service.

3.2.3 Bioinformatics tools and software

Table 3.3 shows software, databases and websites used in this study

3.2.4 Taxonomic classification

All sequence data were trimmed using Trim galore and paired ends were joined using FLASH. Four methods were used for taxonomic classification of datasets: Kraken, MEGAN, One Codex and QIIME.

Kraken was run using the standard Kraken database and visualisation of results was created using KRONA.

BLAST was run using BLASTn, an e value of 0.00001 and the SILVA database. BLAST output was imported to MEGAN, with minimum support of 15 and the 16S percent identity filter on.

FASTA files were uploaded to One Codex and the 'Targeted Loci' database selected.

QIIME was run using the following workflow:

1. add_qiime_labels.py
2. pick_otus.py
3. pick_rep_set.py
4. assign_taxonomy.py
5. make_otu_table.py
6. filter_samples_from_otu_table.py –n 2
7. normalize_table.py
8. summarize_taxa_through_plots.py

.

Table 3.3 Software and websites used in this chapter

Name	Version	Available from:	Reference
BLAST	2.2.26	http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download	Camacho *et al*., 2009
FLASH	1.2.7	https://ccb.jhu.edu/software/FLASH/	Magoc & Salzberg, 2011
KRAKEN	0.10.6	https://ccb.jhu.edu/software/kraken/	Wood & Salzberg, 2014
MEGAN	5.7.0	http://ab.inf.uni-tuebingen.de/software/megan/	Huson *et al*., 2007
Mocrobiota	-	http://mockrobiota.caporasolab.us/	Bokulich *et al*., 2016
Mummer	3.23	http://mummer.sourceforge.net/	Kurtz *et al*., 2004
One Codex		https://www.onecodex.com/	Minot *et al*., 2015
QIIME (MacQIIME)	1	http://www.wernerlab.org/software/macqiime	Caporaso *et al*., 2010
SILVA	128	https://www.arb-silva.de/	Pruesse *et al*., 2007
Trim galore	0.3.3	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/	Krueger, 2015

3.3 Assessing four tools for 16S rRNA taxonomic classification

As previously discussed (section 3.1.2) the ways bioinformatics tools perform taxonomic assignments of 16S rRNA gene sequence data vary and their levels of accuracy have been assessed (Lindgreen *et al*., 2016; Nilikanta *et al.*, 2014; Plummer *et al*., 2015). In order to select an appropriate tool for the analysis of the 16S rRNA gene sequence dataset generated in this chapter, four methods were chosen for further assessment of their accuracy. The methods selected were: Kraken, MEGAN, One Codex and QIIME. These tools were chosen as they involve different levels of user interaction; Kraken and QIIME are run through the command line, MEGAN can be used through both a GUI and the command line whilst One Codex uses a web-based GUI. These four tools also utilise different methods of taxonomic classification, detailed below, and additionally they performed relatively well in the study by Lindgreen *et al.* (2015) (Table 3.2).

Two mock communities were selected from 'Mocrobiota': Mock 12 and Mock 13. Both mock community sequence data sets were generated on the Illumina MiSeq, with the V4 region of the 16S rRNA gene targeted using 515f-806r primers. Mock 12 is composed of 27 bacterial strains, generated in a study by Callahan *et al*. (2016) a number of the taxa are closely related, with some strains having very little difference in their 16S rRNA gene sequences (Table 3.4). Mock 13 is composed of 21 bacterial strains, generated in a study by Kozich *et al.* (2013) (Table 3.5). Mock 12 contained 2 040 485 paired end reads and Mock 13 contained 602 819 paired end reads; after trimming and joining the paired ends using FLASH these read numbers were reduced to 1 878 179 for Mock 12 and 402 483 for Mock 13.

Table 3.4 Bacterial species present in Mock 12

Mock community member	Proportion of mock community (%)	Proportion at genus level (%)
Bacteroides cellulosilyticus DSM 14838	3.7	29.6
Bacteroides eggerthii BEI HM-210	3.7	
Bacteroides fragilis ATCC 23745	3.7	
Bacteroides massiliensis JCM 12982	3.7	
Bacteroides ovatus DSM 1896	3.7	
Bacteroides thetaiotaomicron DSM 2079	3.7	
Bacteroides uniformis DSM 6597	3.7	
Bacteroides vulgatus DSM 1447	3.7	
Barnesiella intestinihominis DSM 21032	3.7	3.7
Clostridium celatum JCM 1394	3.7	18.5
Clostridium cocleatum DSM 1551	3.7	
Clostridium methylpentosum DSM 5476	3.7	
Clostridium phytofermentans ATCC 700394	3.7	
Clostridium xylanovorans DSM 12503	3.7	
Coprococcus comes ATCC	3.7	3.7
Eubacterium rectale DSM 17629	3.7	3.7
Howardella ureilytica DSM 15118	3.7	3.7
Parabacteroides distasonis JCM 13400	3.7	14.8
Parabacteroides distasonis JCM 13401	3.7	
Parabacteroides merdae DSM 19495	3.7	
Parabacteroides sp. D13 BEI HM-77	3.7	
Paraprevotella clara DSM 19731	3.7	3.7
Prevotella buccalis ATCC 35310	3.7	7.4
Prevotella copri DSM 18205	3.7	
Roseburia intestinalis DSM 14610	3.7	7.4
Roseburia inulinivorans DSM 16841	3.7	
Ruminococcus gnavus ATCC 29149	3.7	3.7

Table 3.5 Bacterial species present in Mock 13

Mock community member	Proportion of mock community (%)	Proportion at genus level (%)
Acinetobacter baumannii ATCC 17978	4.8	4.8
Actinomyces odontolyticus ATCC 17982	4.8	4.8
Bacillus cereus ATCC 10987	4.8	4.8
Bacteroides vulgatus ATCC 8482	4.8	4.8
Clostridium beijerinckii ATCC 51743	4.8	4.8
Deinococcus radiodurans DSM 20539_	4.8	4.8
Enterococcus faecalis ATCC 47077	4.8	4.8
Escherichia coli ATCC 700926	4.8	4.8
Helicobacter pylori ATCC 700392	4.8	4.8
Lactobacillus gasseri DSM 20243	4.8	4.8
Listeria monocytogenes ATCC BAA-679	4.8	4.8
Neisseria meningitidis ATCC BAA-335	4.8	4.8
Porphyromonas gingivalis ATCC 33277	4.8	4.8
Propionibacterium acnes DSM16379	4.8	4.8
Pseudomonas aeruginosa ATCC 47085	4.8	4.8
Rhodobacter sphaeroides ATCC 17023	4.8	4.8
Staphylococcus aureus ATCC BAA-1718	4.8	9.6
Staphylococcus epidermidis ATCC 12228	4.8	
Streptococcus agalactiae ATCC BAA-611	4.8	14.4
Streptococcus mutans ATCC 700610	4.8	
Streptococcus pneumoniae ATCC BAA-334	4.8	

3.3.1 Classification using Kraken

The taxonomic classifier Kraken analyses all sequences in a data set and is run through the command line. Kraken claims to have overcome the problem of classifiers either being accurate but slow or fast but less sensitive by utilizing exact alignments of k-mers and a novel classification algorithm (Wood & Salzberg, 2014). The standard kraken database is composed of records consisting of a *k*-mer and the lowest common ancestor (LCA) of all organisms whose genomes contain that *k*-mer. Query sequences are broken into k-mers and aligned to those in the database, the associated set of LCA taxa are then used to determine an appropriate label for the sequence. Sequences with no k-mers in the database remain unclassified (Figure 3.2).

Kraken classified 99.9 % of both mock data sets used in this study (Figures 3.3 and 3.4). Like the other classifiers assessed (see below), Kraken did not produce very accurate results for mock 12, with a high proportion of reads (90 %) being identified as *Bacteroides* and eight out of the 11 genera having either no reads assigned to them or less than 1 % of the reads assigned to them (Table 3.6). However, Kraken produced more accurate results for mock 13; fourteen out of eighteen genera from mock 13 were correctly identified with at least 1 % of classified reads assigned to them. One false positive was assigned from mock 13 with *Kribbella flavida* identified as having 3 % of reads assigned to it. A positive feature of Kraken is the detailed output it provides; through the use of the Kraken-translate script the user is able to inspect which reads have been assigned to certain taxa. This allowed for the reads assigned to *Kribbella* to be extracted and used in a BLAST query against the NCBI 16S rRNA database. The results from this BLAST search showed that the sequences Kraken had assigned to *Kribbella* were actually *Actinomyces odontolyticus*, a species which is present in mock 13 but which was not classified by Kraken. Both *Kribbella* and *Actinomyces* are from the order Actinomycetales. This demonstrates an error in Krakens assignment at the order level, although this was the only false positive produced for this dataset.

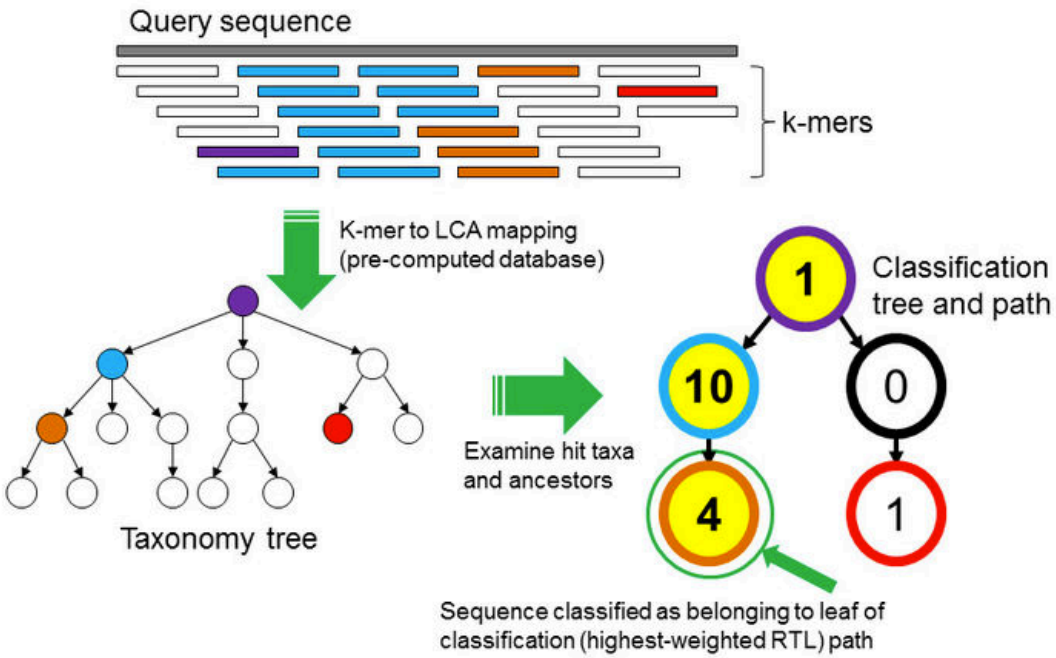


Figure 3.2 The Kraken classification algorithm. Sequences are broken into k-mers which are then mapped to the lowest common ancestor (LCA) of genomes in the kraken database containing that k-mer. A classification tree is formed from taxa associated with the sequences k-mers. Nodes in the classification tree have a weight equal to the number of k-mers in the sequence associated with the node's taxon, the sum of these weights gives a score to each root-to-leaf (RTL) path. The maximal RTL path in the classification tree forms the classification path (nodes highlighted in yellow). The leaf of this classification path (the orange, leftmost leaf in the classification tree) is the classification used for the query sequence. (Wood and Salzberg, 2014).

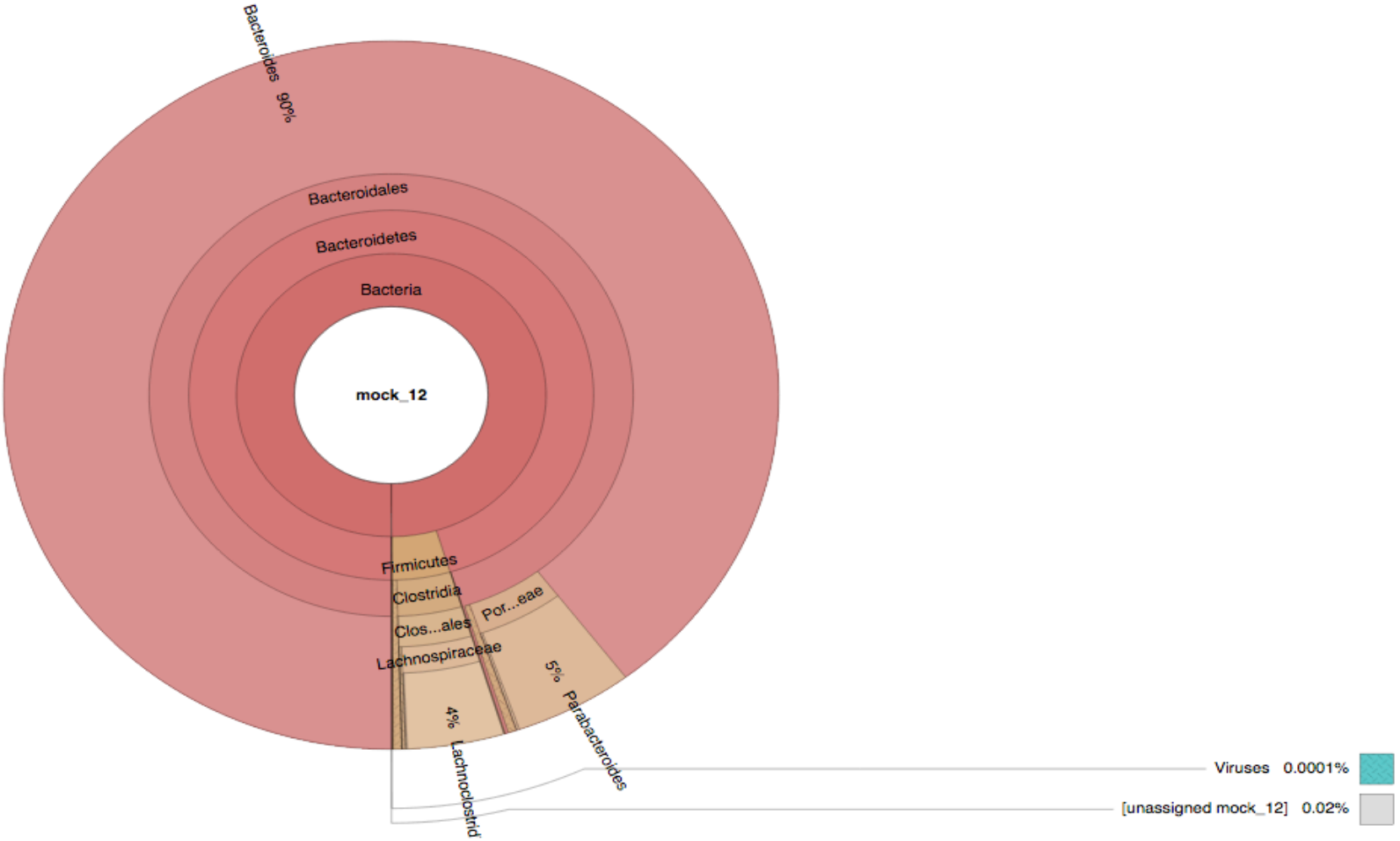


Figure 3.3 Taxonomic classification (to lowest possible taxonomic level, down to genus) of Mock 12 by Kraken.

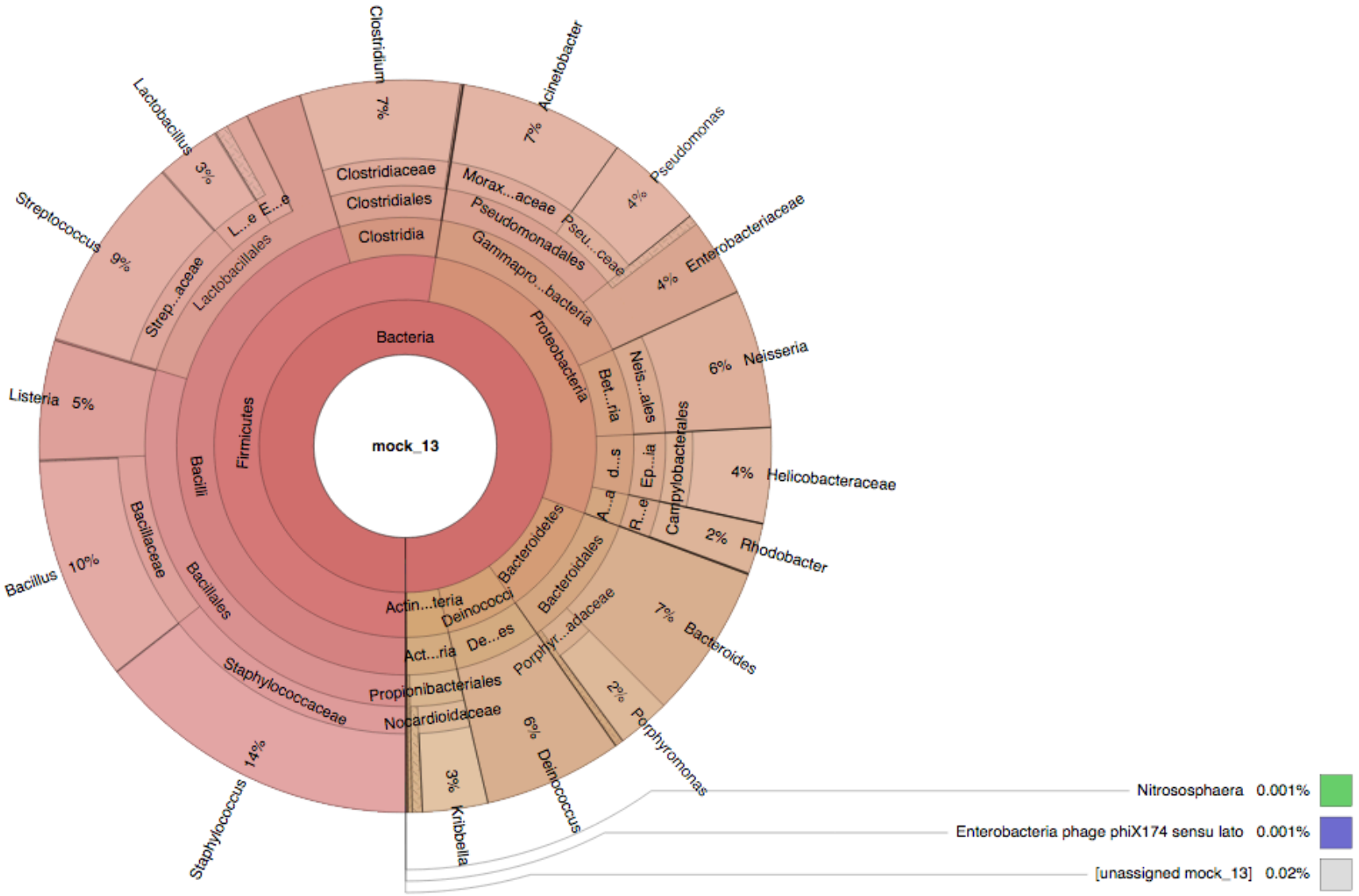


Figure 3.4 Taxonomic classification (to lowest possible taxonomic level, down to genus) of Mock 13 by Kraken.

3.3.2 Classification using MEGAN

The taxonomic classifier MEGAN analyses all sequences in a data set and can be run through a GUI or via the command line. The main pre-processing step of MEGAN is to use BLAST (or another alignment tool) to compare the sequences to known sequences in a database. The BLAST output is then imported to MEGAN, which applies a lowest common ancestor algorithm to classified sequences, summarises and orders the results and provides an interactive tool for exploring the BLAST results (Figure 3.5). BLAST is a widely used and accurate tool, however when dealing with large datsets it is also time consuming and computationally intensive, producing large files as output. Within this study the SILVA database was used as the database in the BLAST alignment and MEGAN was used to visualise output. Due to the large number of sequences in the dataset mock 12, computational problems arose during the BLAST process and ultimately it has not been possible to analyse this data set. However, mock 13 was successfully analysed Figure 3.6 shows the MEGAN output for mock 13.

When using a threshold of 1 % for minimum support MEGAN correctly classified 10 out of the 18 genera present in the mock 13 dataset and produced no false positives. As with all classifiers discussed here, the proportion classified as each genus varied, although it was close to the expected value for *Actinomyces*, *Deinococcus*, *Neisseria* and *Pseudomonas* with 3.5, 3.0, 3.3 and 5.0 percent of reads assigned to these genera respectively (Expected value = 3.8 %).

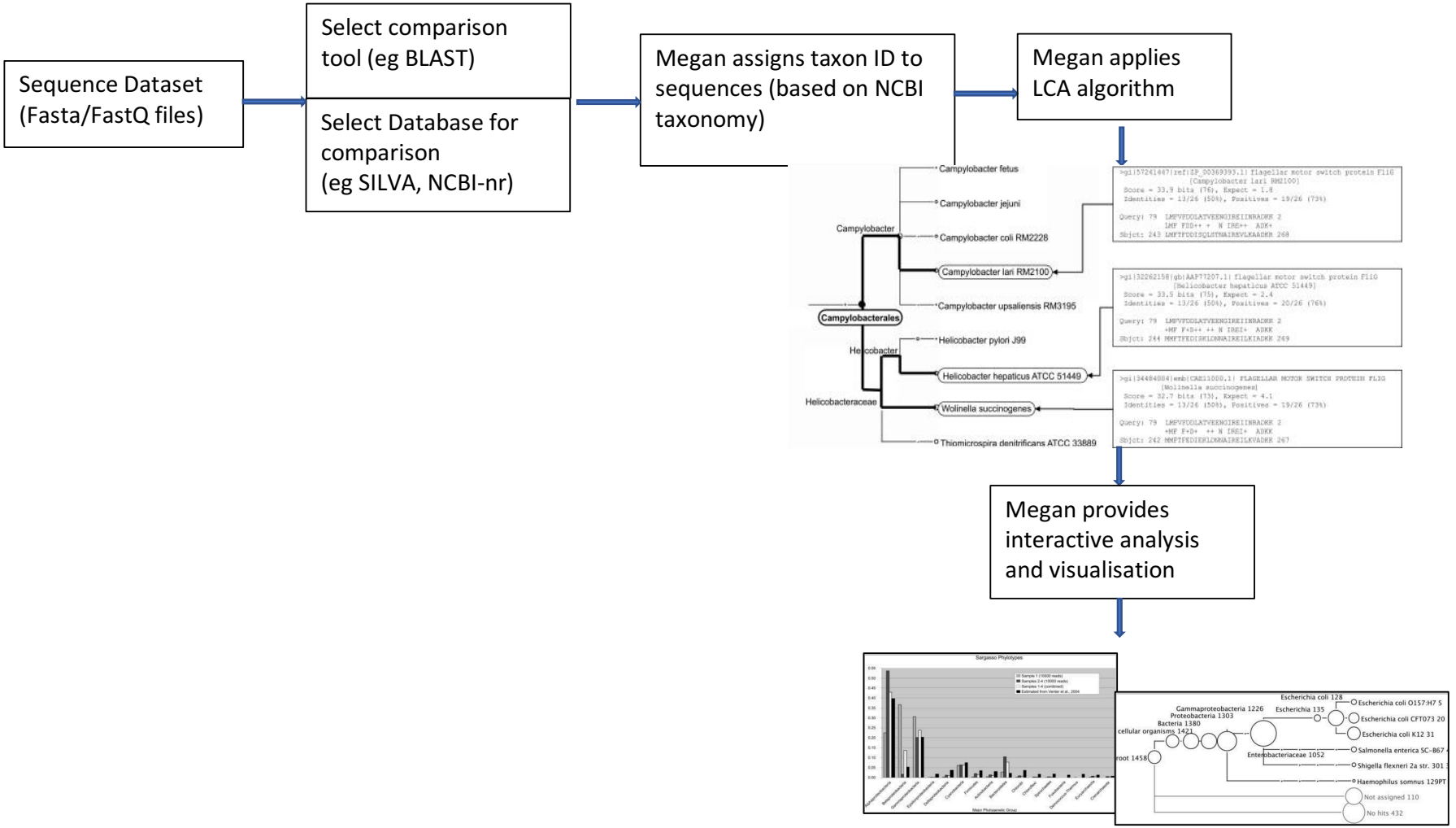


Figure 3.5 Basic workflow for MEGAN: A comparison tool is used to compare unknown sequences against known sequences in a database, output is uploaded to MEGAN which assigns taxon ID based on NCBI taxonomy, MEGAN then applies a Lowest Common Ancestor (LCA) assignment algorithm to the dataset and displays the results in a format which the user can interact with.

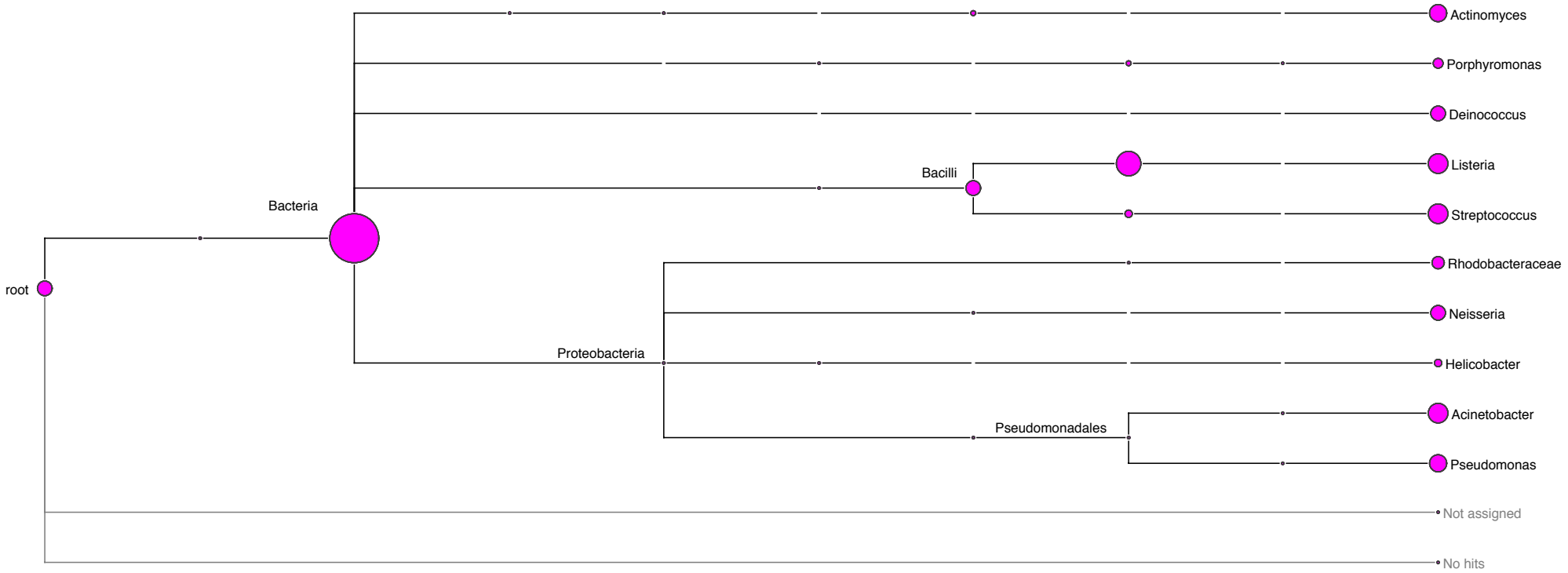


Figure 3.6 Taxonomic classification (to lowest possible taxonomic level, down to genus) of Mock 13 using MEGAN

3.3.3 Classification using One Codex

The taxonomic classifier One Codex analyses all sequences in a genomic or 16S rRNA sequence data set and is run through a web based GUI. One Codex breaks reads into *k*-mers of length 31bp. These *k*-mers are then compared against one of three user-selected databases; the available databases are: The RefSeq database, 8120 genomes from bacteria, viruses, fungi and archaea; The One Codex expanded database, 30825 genomes from bacteria, viruses, fungi, archaea and protozoa; The target loci database, 247646 records covering 5S, 16S, 23S, *gyrB*, *rpoB*, 18S, 28S and ITS genes. Reads are then summarized as "*k*-mer hit chains" describing the complete set of taxonomically informative *k*-mers from within each read. Each read is then assigned to a taxon based on the LCA of all *k*-mers in the "hit chain". One Codex allows for the user to select the database used and to select which nodes to display, based on a threshold for number of reads assigned when displaying results as a taxonomic chart (1 % was used in this study). However, no other interaction between the user and the software is possible.

As with Kraken, One Codex was unable to accurately analyse the mock 16S rRNA sequence dataset 12, with high levels of false negatives and very inaccurate abundance estimations (Figure 3.7, Table 3.6). For mock 13 it produced no false positives and four false negatives (Figures 3.8 and Table 3.7).

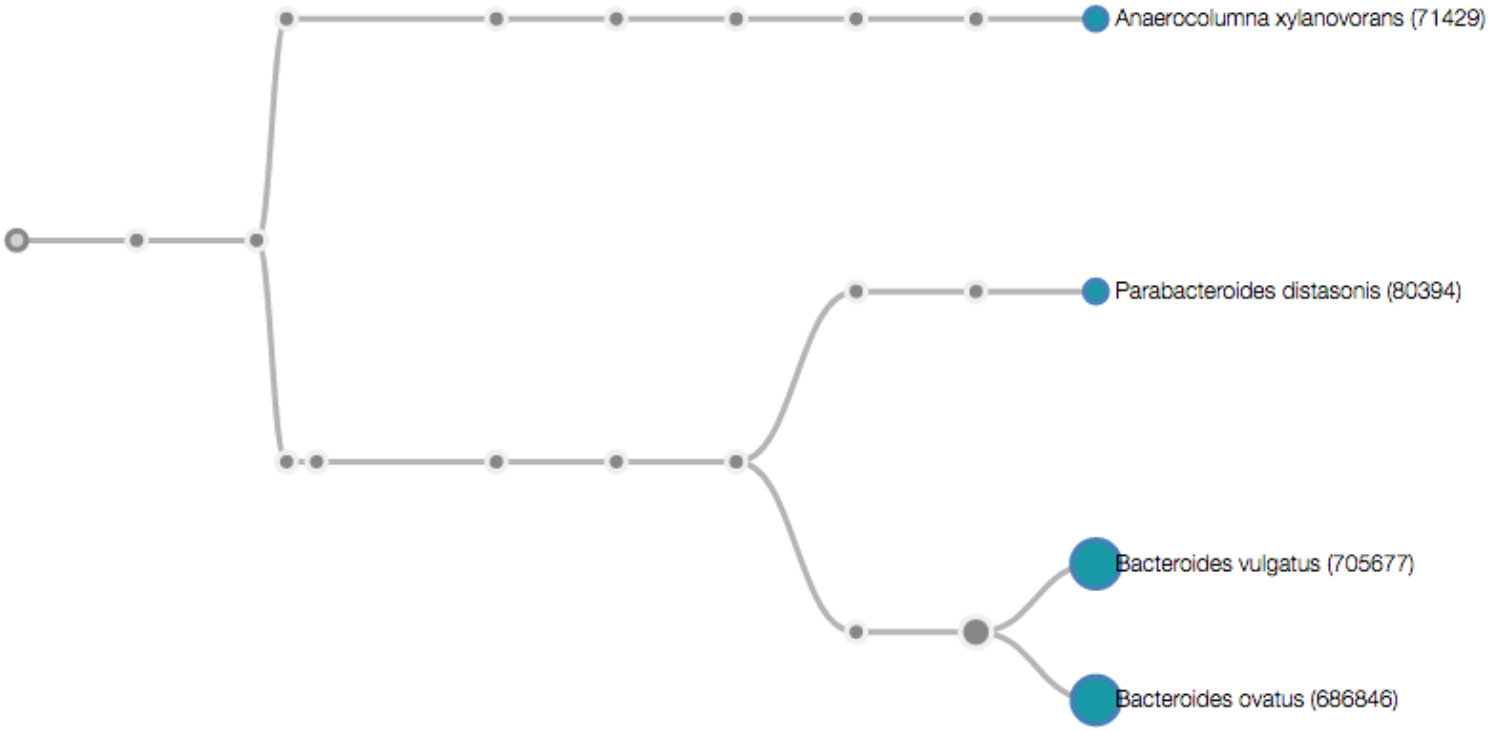


Figure 3.7 Classification of mock 12 by One codex

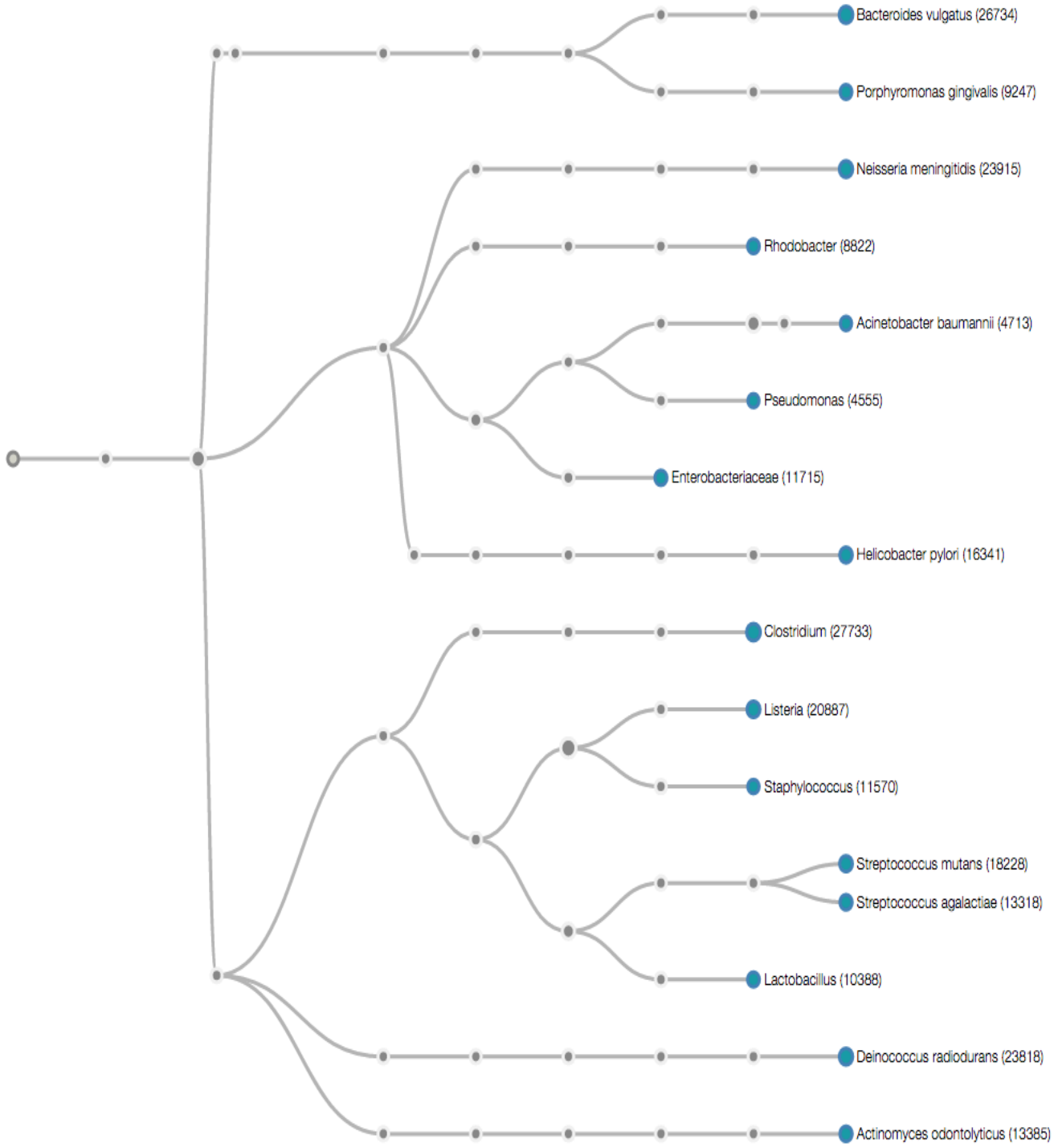


Figure 3.8 Classification of mock 13 by One codex

3.3.4 Classification using QIIME

The taxonomic classifier Quantitative Insights into Microbial Ecology (QIIME) analyses selected marker genes from a community and is run through the command line. The QIIME package is composed of a number of Python scripts which can be selected by the user to create an appropriate pipeline for the analysis of 16S rRNA datasets. The modular nature of the QIIME package allows for the user to have full control over parameters used in each step of the analysis as well as the option to select only a specific set of scripts appropriate to the dataset being looked at. The set of QIIME scripts used for the analysis of the mock datasets in this study, as well as for subsequent analysis of the 16S rRNA datasets of bacteria associated with *B. braunii* (Sections 3.4 and 3.5) is shown in Figure 3.9.

Like the previously discussed methods, QIIME was not able to classify the majority of species present in mock dataset 12 (Figure 3.10). However, out of all the methods looked at, QIIME was most successful at classifying mock dataset 13, with no false negatives and only one false positive. The accuracy of the abundance estimations for each species was variable (Table 3.7)

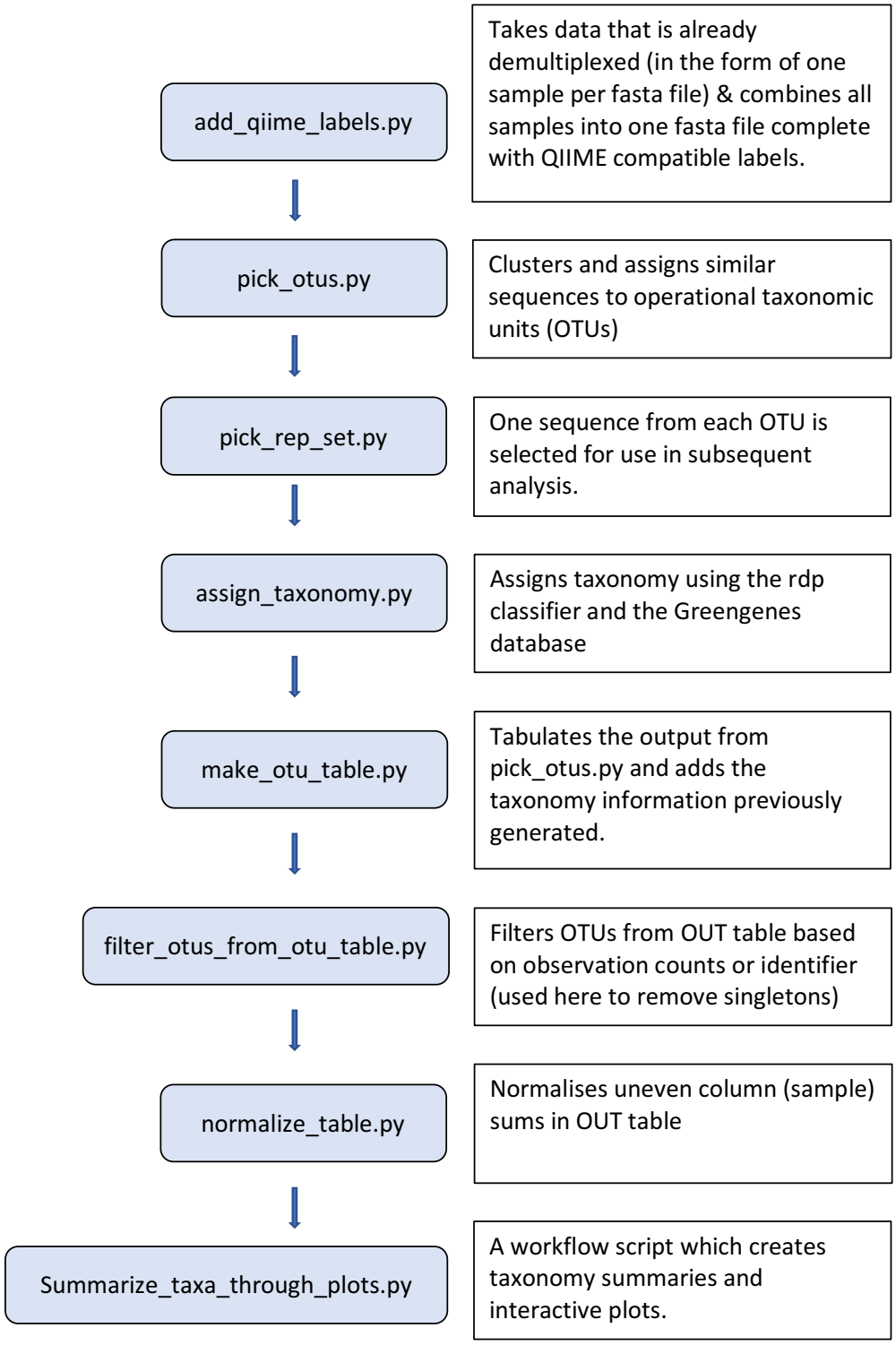


Figure 3.9 QIIME (version one) scripts used in this study and their function.

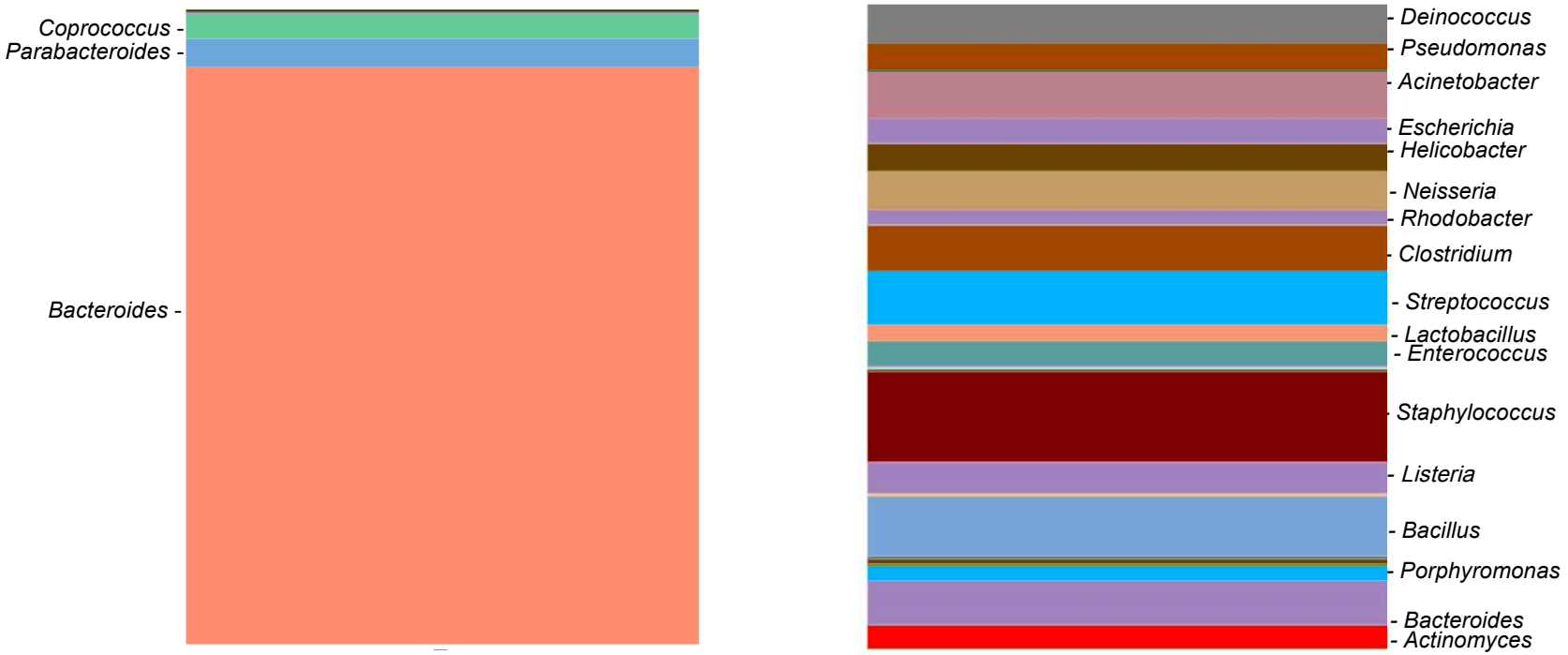


Figure 3.10 Plots generated by QIIME showing taxonomic classifications for mock communities 12 and 13.

3.3.5 Assessment of the four methods used

None of the tools used were able to give accurate abundance estimations for all species present; this is consistent with the study by Lindgreen *et al*. (2016) which assessed the accuracy of tools for classifying shotgun metagenomic samples (Section 3.1.2). However, it should be noted that a number of reads from the original data sets were lost following trimming and joining (Table 3.9) and therefore abundance levels would vary from the levels in Tables 3.4 and 3.5. However, loss of reads due to quality control cannot explain the large variation in results demonstrated between each of the methods used and although all tools classified at least 98.2 % of reads, the number assigned to each genus was largely inconsistent. For example, whilst QIIME and Kraken classified 14 % of reads from Mock 13 to the genus *Staphylococcus* (Expected amount: 9.6 %) MEGAN did not classify any reads to this genus and One Codex classified 2.9 %. These differences must be attributed to the classification system used by the different tools.

All of the methods used performed extremely poorly in classifying mock data set 12, highlighting the difficulties of differentiating between highly similar 16S rRNA gene sequences. However, as with mock data set 13, mock 12 is largely composed of bacteria of clinical interest and within environmental samples there is likely to be a greater variation in species present.

Overall QIIME performed best in terms of fewest false negatives followed by One codex, Kraken and MEGAN (for mock dataset 13). When using a minimum threshold of 1% (of classified reads) false negatives outweigh false positives. Therefore, by using this threshold in future studies we can have confidence in positive results reported but it is possible species in low abundance may be overlooked. QIIME and Kraken were both faster than MEGAN and allowed for more user interaction than One Codex. Kraken had useful tools for generating reports and extraction of specific reads based on their taxonomic classification is easier with Kraken compared to QIIME. Kraken is also relatively quick to run and useful for quick, initial analysis of the complexity of a sample. Therefore, either QIIME or Kraken will be used for analysis in subsequent sections of this thesis.

Table 3.6 Mock Community 12 composition and classifications by QIIME, Kraken and One Codex.

Mock community member	Proportion in mock community (%)	Proportion classified to genus level using QIIME (%)	Proportion classified to genus level using Kraken (%)	Proportion classified to genus level using One Codex (%)
Bacteroides cellulosilyticus DSM 14838	3.7	90.19	90	90.31
Bacteroides eggerthii BEI HM-210	3.7			
Bacteroides fragilis ATCC 23745	3.7			
Bacteroides massiliensis JCM 12982	3.7			
Bacteroides ovatus DSM 1896	3.7			
Bacteroides thetaiotaomicron DSM 2079	3.7			
Bacteroides uniformis DSM 6597	3.7			
Bacteroides vulgatus DSM 1447	3.7			
Barnesiella intestinihominis DSM 21032	3.7	Not classified	Not classified	Not classified
Clostridium celatum JCM 1394	3.7	Not classified	4	0.04
Clostridium cocleatum DSM 1551	3.7			
Clostridium methylpentosum DSM 5476	3.7			
Clostridium phytofermentans ATCC 700394	3.7			
Clostridium xylanovorans DSM 12503	3.7			
Coprococcus comes ATCC	3.7	3.68	0.002	0.29
Eubacterium rectale DSM 17629	3.7	Not classified	0.04	Not classified
Howardella ureilytica DSM 15118	3.7	Not classified	Not classified	Not classified
Parabacteroides distasonis JCM 13400	3.7	4.77	5.0	4.8
Parabacteroides distasonis JCM 13401	3.7			
Parabacteroides merdae DSM 19495	3.7			
Parabacteroides sp. D13 BEI HM-77	3.7			
Paraprevotella clara DSM 19731	3.7	Not classified	Not classified	Not classified

Prevotella buccalis ATCC 35310	3.7	Not classified	0.3	Not classified
Prevotella copri DSM 18205	3.7			
Roseburia intestinalis DSM 14610	3.7	Not classified	0.005	0.04
Roseburia inulinivorans DSM 16841	3.7			
Ruminococcus gnavus ATCC 29149	3.7	Not classified	0.0006	Not classified

Table 3.7 Mock Community 13 composition and classification by QIIME, Kraken, MEGAN and One Codex.

Mock community member	Proportion in mock community (%)	Proportion classified to genus level using QIIME (%)	Proportion classified to genus level using Kraken (%)	Proportion classified to genus level using MEGAN(%)	Proportion classified to genus level using One Codex (%)
Acinetobacter baumannii ATCC 17978	4.8	7.3	7	6.0	7
Actinomyces odontolyticus ATCC 17982	4.8	3.5	Not Classified	4.5	3.5
Bacillus cereus ATCC 10987	4.8	9.3	10	0.04	0.9
Bacteroides vulgatus ATCC 8482	4.8	6.8	7	0.03	6.8
Clostridium beijerinckii ATCC 51743	4.8	7.0	7	0.03	7.0
Deinococcus radiodurans DSM 20539	4.8	6.1	6	4.0	6.0
Enterococcus faecalis ATCC 47077	4.8	3.8	0.5	0.03	0.8
Escherichia coli ATCC 700926	4.8	3.7	0.9	Not classified	0.1
Helicobacter pylori ATCC 700392	4.8	4.2	4 *	1.3	4.2
Lactobacillus gasseri DSM 20243	4.8	2.5	3	0.8	2.6
Listeria monocytogenes ATCC BAA-679	4.8	4.7	5	6.9	5.3
Neisseria meningitidis ATCC BAA-335	4.8	6.0	6	4.3	6.1
Porphyromonas gingivalis ATCC 33277	4.8	2.2	2	2.1	2.4
Propionibacterium acnes DSM16379	4.8	0.1	0.8	0.6	Not Classified
Pseudomonas aeruginosa ATCC 47085	4.8	4.1	4	5.0	1.2
Rhodobacter sphaeroides ATCC 17023	4.8	2.2	2	3.1 *	2.2
Staphylococcus aureus ATCC BAA-1718	4.8	13.8	14	Not classified	2.9
Staphylococcus epidermidis ATCC 12228	4.8				
Streptococcus agalactiae ATCC BAA-611	4.8	8.3	9	6.9	8.6
Streptococcus mutans ATCC 700610	4.8				
Streptococcus pneumoniae ATCC BAA-334	4.8				

* Only classified to family level

Table 3.8 Numbers of false positives, false negatives and reads classified for Mock dataset 13.

Method	**Number of genera correctly assigned (out of 18) ***	**Number of false positives ***	**Number of false negatives ***	**% of reads classified**
Kraken	14	1	4	99.9
MEGAN	10	0	8	99.9
One Codex	14	0	4	98.2
QIIME	17	0	1	100

* Only genera with >1% of reads classified to it are included here.

3.4 Analysis of different 16S rRNA variable regions results in different taxonomic distribution

The bacterial 16S rRNA gene is around 1.5 kb long and contains conserved regions along with nine hypervariable regions. Currently, selected regions are targeted for sequencing, with which variable region is the best for taxonomic classification being a topic of debate (Chakravorty *et al*. 2007; Kumar et al., 2011). Previous studies have found that the 16S variable region targeted can greatly affect the outcome of taxonomic classification, with differences in species distribution present when different variable regions are used (Liu *et al*. 2008; Claesson *et al*., 2010). In this chapter, the V1-V2, V3-V4 and V5-V8 variable regions of the 16s rRNA genes from bacteria found in loose and close association (Datasets A and B respectively) with *B. braunii* have been sequenced. Table 3.9 shows statistics for each of these sequencing data sets.

3.4.1 Taxonomic classification using QIIME shows different results depending on variable region used.

The taxonomic classifier QIIME was selected for analysis of the bacterial communities in loose and close association with *B. braunii* (Data sets A and B respectively). QIIME produced no false positives and only one false negative in the mock data set analysed previously (Section 3.3) and allows the user to tailor the workflow to suit the needs of their own specific study. QIIME is also designed to analyse amplicon data making it a good option for this study. Table 3.10 shows the percentage of reads, from loose and close association with *B. braunii*, classified using QIIME.

When looking at QIIME results at the genus level, large variation is seen in the taxonomic distribution for both data sets when using different variable regions (Figures 3.11 - 3.14). However, some of these differences can be attributed to differences in the lowest taxonomic level that it has been possible to classify to. Within data set A the V5-V8 region has been ineffective at classifying a large proportion of reads to the genus level, with 43.3% classified as Gammaproteobacteria (class) and a further 14.4% classified as bacteria

(kingdom). The percentage of reads classified to the order level or lower for Set A was 81.8 %, 90.9 %, and 17.9 % for the V1-V2, V3-V4 and V5-V8 variable regions respectively. The V5-V8 variable region was more effective for classifying to a lower taxonomic level when looking at dataset B with 66.3 % classified to order level or lower. The V1-V2 and V3-V4 variable regions classified 89.6 % and 93.2 % of reads to the order level or lower respectively.

Table 3.9 Sequence statistics for each variable region targeted, from Set A (loose association with *B. braunii*) and Set B (close association with *B. braunii*)

Data set	Total number of reads (in fastq files)	Number of reads assigned to V1_V2 *	Number of reads assigned to V3_V4 *	Number of reads assigned to V5_V8 *	Total number of reads assigned to one of three variable regions.
Set A	3 183 489	1 409 119 (44.3 %)	868 094 (27.3 %)	219 881 (6.9 %)	2 497 094 (78.4 %)
Set B	1 668 445	459 488 (27.5 %)	195 356 (11.7 %)	969 819 (58.1 %)	1 624 663 (97.4 %)

*After trimming, FLASH and assignment to one of three variable regions determined on the presence of a forward primer.

Table 3.10 Percentage of reads classified from each variable region targeted, from Set A (loose association with *B. braunii*) and Set B (close association with *B. braunii*)

Data set	**V1-V2 region** % of reads classified using QIIME	**V3-V4 region** % of reads classified using QIIME	**V5-V8 region** % of reads classified using QIIME	**V1-V2 region** % of reads classified to order level or lower	**V3-V4 region** % of reads classified to order level or lower	**V5-V8 region** % of reads classified to order level or lower
Set A	93.9	96.0	95.0	81.0	90.9	17.9
Set B	95.8	98.1	96.7	89.6	93.2	66.3

V1-V2 region V3-V4 region V5-V8 region

Figure 3.11 Plots generated using QIIME demonstrate differences in taxonomic distribution for dataset A, depending on which variable region is used. Taxa with >1% of reads assigned are labelled. The lowest taxonomic rank it has been possible to classify to is shown (p=phylum, c=class, o-order, f=family, g=genus).

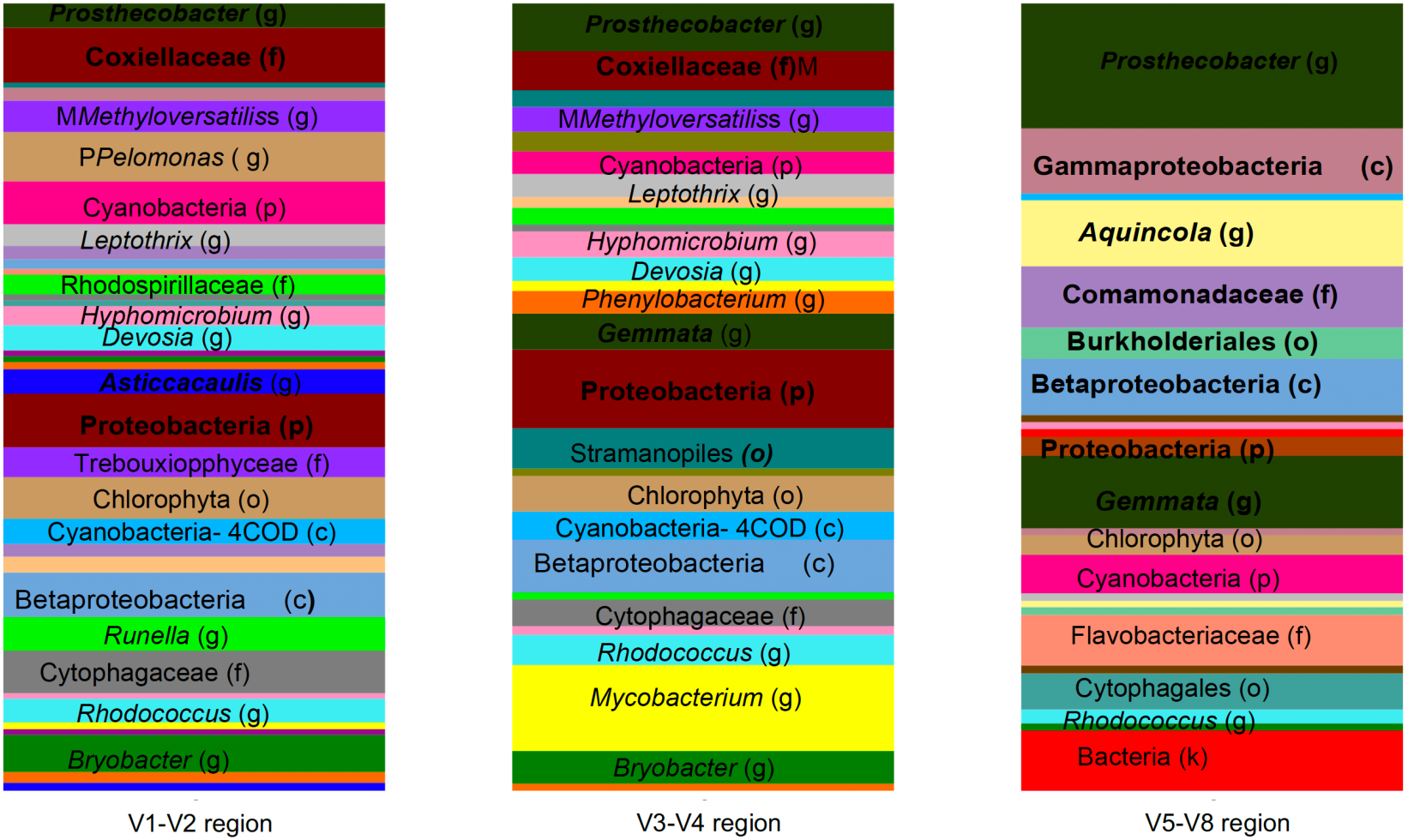


Figure 3.12 Plots generated using QIIME demonstrate differences in taxonomic distribution for dataset B, depending on which variable region is used. Taxa with >1% of reads assigned are labelled. The lowest taxonomic rank it has been possible to classify to is shown (p=phylum, c=class, o-order, f=family, g=genus).

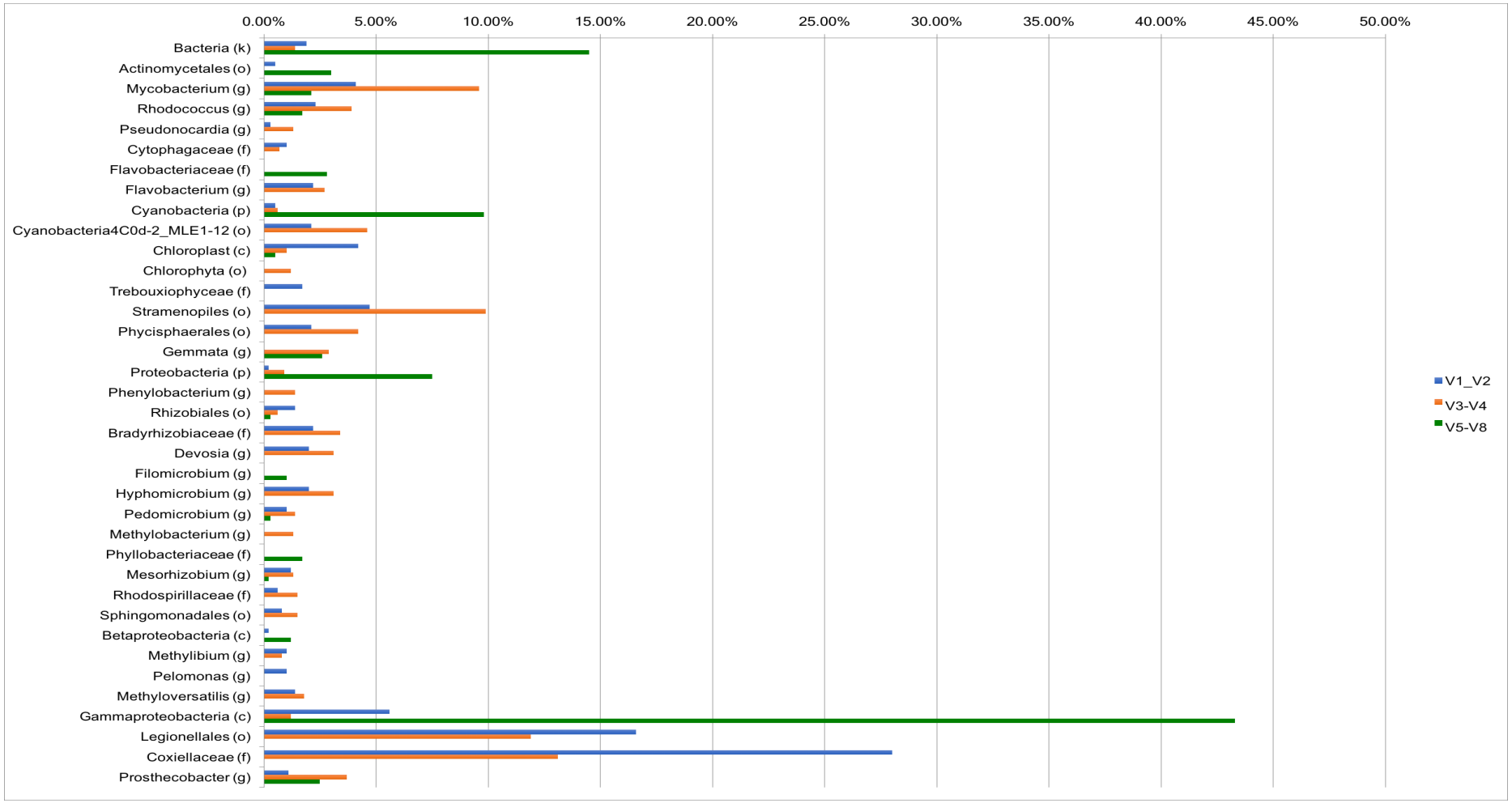


Figure 3.13 QIIME taxonomy classifications (Set A)

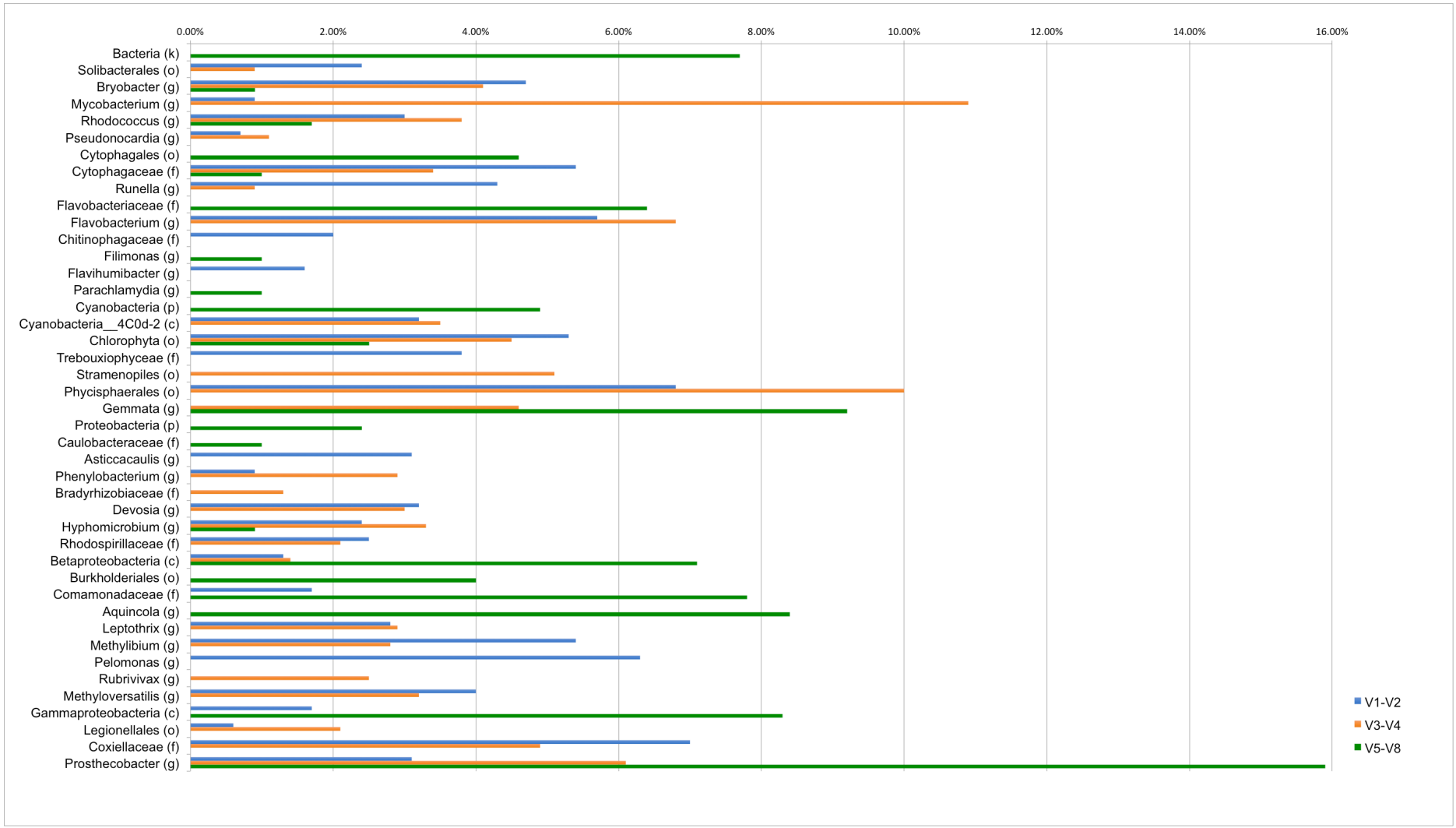


Figure 3.14 QIIME taxonomy classifications (Set B)

3.5 A diverse range of bacteria are present in both close and loose assocuiation with *B. braunii*

Bacteria found living alongside *B. braunii* can be placed into one of two categories: those living in close association (forming biofilms on the surface of the microalgae) and those living in loose association (planktonically in the water column) (Rivas *et al*., 2010). Differences in ways in which the two bacterial populations may be interacting with *B. braunii* have previously been observed, with Rivas *et al.* observing thst quorum sensing signals were present in two out of eight bacteria cultured from close association with *B. braunii*, whilst no quorum sensing signals were found in bacteria cultured from loose association. No studies have been carried out to determine the taxonomy of all bacteria present in consortium with *B. braunii* in both close and loose association. Determining which bacteria are present may reveal the presence of bacteria known to interact with eukaryotes and contribute to further understanding of optimal growth conditions for *B. braunii.*

3.5.1 The bacterial populations in close and loose association with B. braunii

QIIME was used to assign taxonomic classification to the 16S rDNA data set to the phylum, class and genus levels (Figures 3.15 - 3.17), although it is worth noting that not all data was classified as low as genus level; Figure 3.17 demonstrates the lowest taxonomic classification achieved. The following sections detail the phyla present and further discuss the lower taxonomic classifications where information is available. All percentages discussed are of the total number of reads classified for each sample.

One of the aims of this chapter was to determine if bacteria isolated from *B. braunii* in chapter 3 are present in loose or close association and if they are representative of abundant taxa or not. Bacteria identified in chapter 3 are: *Achromobacter piechaudii* GCS2, *Agrobacterium* sp. SUL3, *Microbacterium* sp. GCS4, *Shinella* sp. GWS1and *Shinella* sp. SUS2. Three of the four genera identified in chapter 3 are Proteobacteria

(*Achromobacter, Agrobacterium* and *Shinella*) and one is from the phylum Actinobacteriam (*Microbacterium*), these bacteria will therefore also be discussed in the following sections where relevant.

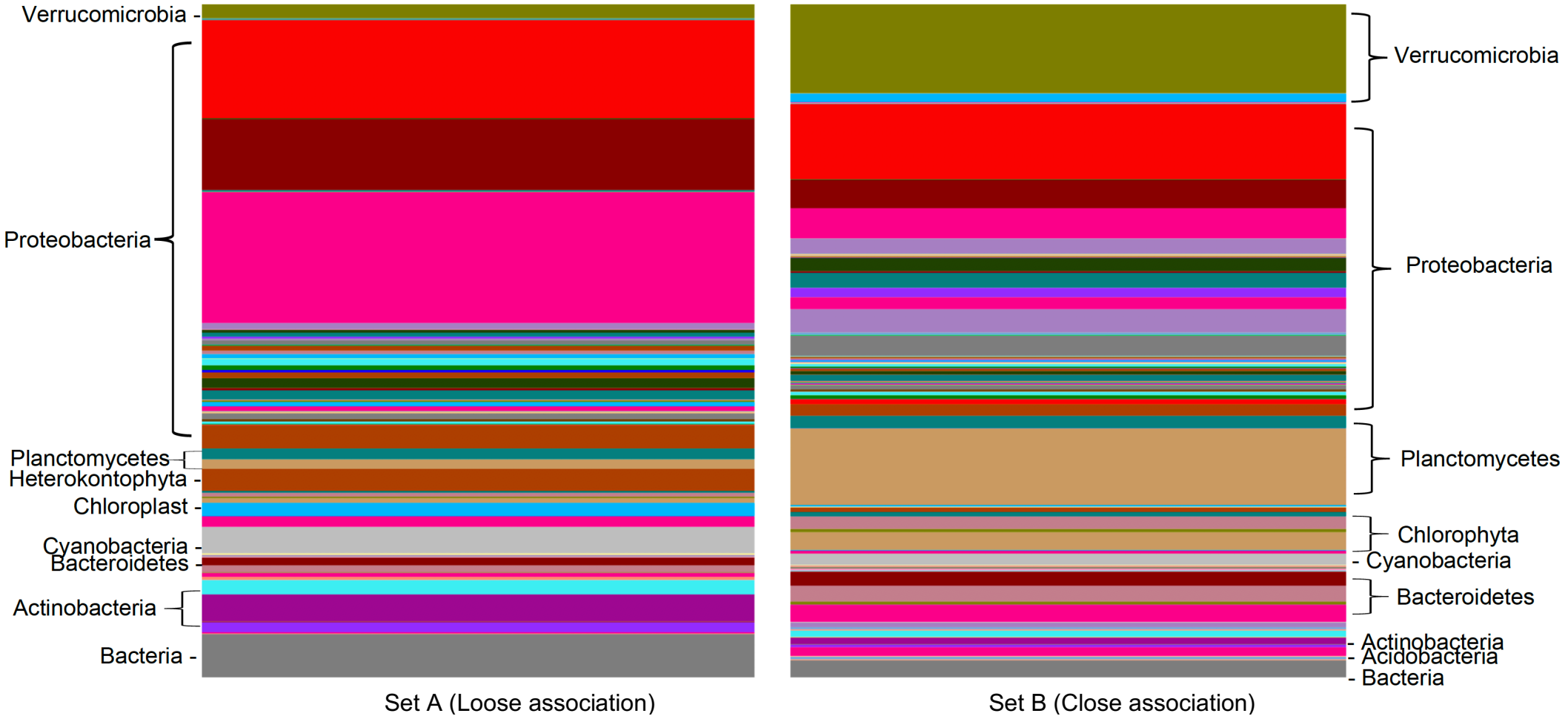


Figure 3.15 Phylum level taxonomic information for data sets A and B assigned by QIIME

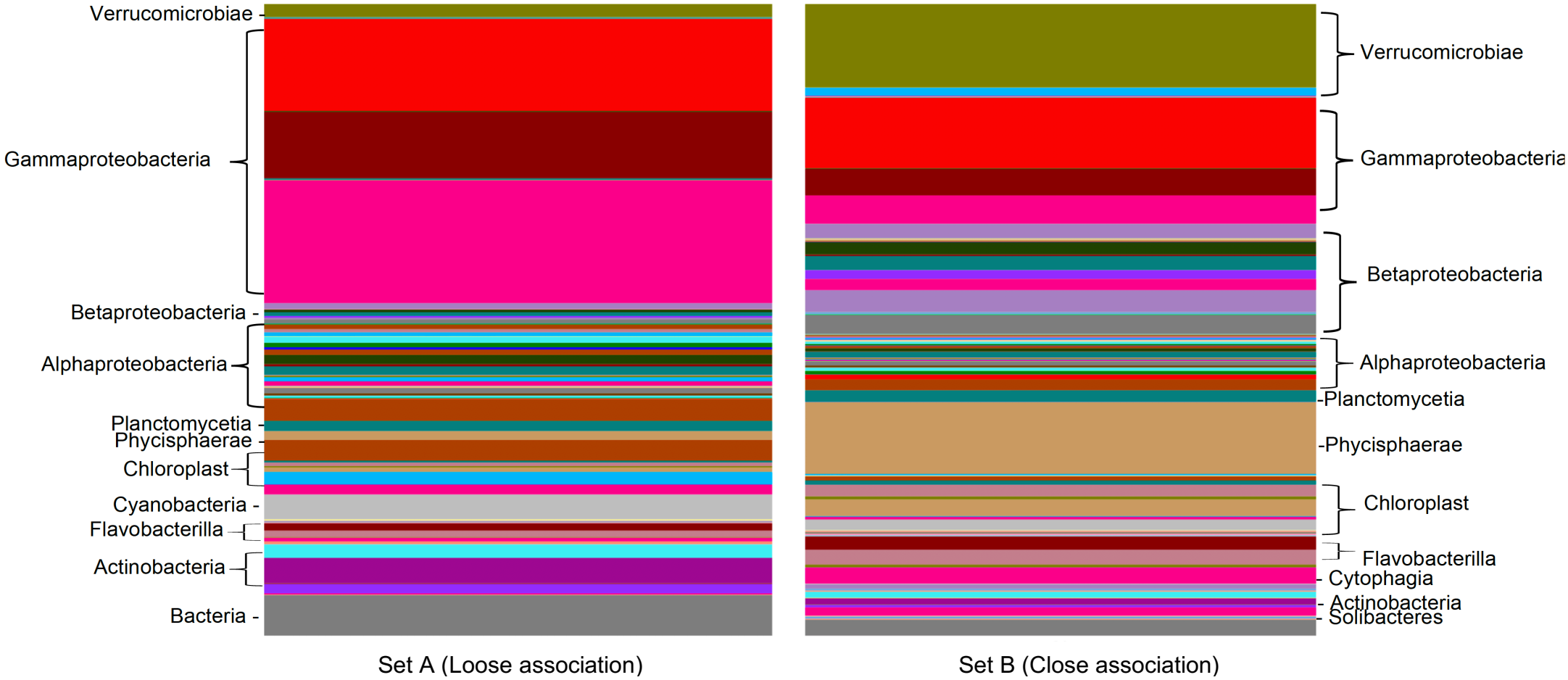


Figure 3.16 Class level taxonomic information for data sets A and B assigned by QIIME.

Figure 3.17 Taxonomic classification for data sets A and B. The lowest taxonomic classification achieved by QIIME is labelled (k = kingdom, p = phylum, c= class, o = order, f = family, g = genus).

Acidobacteria

A small proportion of reads were classified as Acidobacteria (2 % of sample B, <1 % of sample A). Acidobacteria have been found in a diverse range of habitats, including soil, freshwater, wastewater, sewage sludge and hot springs (Quaiser *et al*., 2003). Acidobacteria have been described as one of the most widespread and abundant phyla on the planet, however, difficulties in culturing them have led to a lack of understanding into their role in the numerous ecosystems the inhabit (Kielak *et al*., 2016). There has only been one study linking Acidobacteria to algae; members of the phylum were found to be living in association with the coccolithophorid algae *Emiliania huxleyi* and *Coccolithus pelagicus f. braarudii* (Green *et al.,* 2015). What their potential interactions with both the coccolithophorids and with *B. braunii* are remain unclear, but the discovery of Acidobacteria in this study further aids in the understanding of their distribution.

Actinobacteria

The phylum Actinobacteria is assigned to 8.4 % and 2.7 % of samples A and B respectively. Reads have been assigned to two genera for both samples: *Mycobacterium* and *Rhodococcus*. The genus *Mycobacterium* is large, containing 186 species and 13 subspecies (according to “List of prokaryotic names with standing in nomenclature”, accessed Nov 2017) including well known obligate parasites the *M. tuberculosis* complex and *M. leprae*, the causes of tuberculosis and leprosy respectively. Mycobacteria which do not cause tuberculosis or leprosy are known as Nontuberculous mycobacteria (NTM). NTMs are disinfectant, heavy metal and antibiotic resistant and are found throughout the environment, in water and soil (Falkinham, 2013). Of interest to this study is the fact that NTMs are capable of degrading hydrocarbons, and although studies have typically shown these to be polycyclic aromatic hydrocarbons, there have been instances of NTMs that are able to utilise a wider spectrum of hydrocarbons (Falkinham, 2013; Ferreira, *et al*., 2007; Solano-Serena *et al*., 2000). Due to the obligate parasitic nature of *M. tuberculosis* and *M. leprae* it is most likely that the *Mycobacterium* present in the *B. braunii* consortium is a NTM, and it is likely to be utilising the hydrocarbons that are present.

The second genus that reads from Actinobacteria have been assigned to is *Rhodococcus.* Like *Mycobacterium*, *Rhodococcus* is found in a diverse range of habitats and can withstand harsh environmental conditions (Kuyukina & Ivshina, 2010). Additionally, like *Mycobacterium*, numerous strains of *Rhodococcus* have been found to degrade a range of hydrocarbons (Ruberto *et al*., 20015).

Microbacterium sp. GCS4, a bacterium isolated from *B. braunii* in chapter three, is also from the phylum Actinobacteria. The genus *Microbacterium* has not had any reads classified to it using QIIME nor has the family Microbacteriaceae to which it belongs. However, for sample A and sample B, the order Actinomycetales has 1.5 % and 0.4 % of reads assigned to it which have not been able to be classified to a lower taxonomic level respectively. If *Microbacterium* is present in the bacterial consortia studied within this chapter it must be part of this 1.5 / 0.4 %. It can therefore be concluded that if present, *Microbacterium* is not present in large numbers and is more likely to be found in loose association than close association with *B. braunii*.

Bacteroidetes

3.3 % of sample A and 9.7 % of sample B were assigned to the phylum Bacteroidetes. Both samples had the majority of reads from Bacteroidetes further assigned to the family Flavobacteriaceae and the genus *Flavobacterium.* Consisting of 130 recognised species, the genus *Flavobacterium* is large and found in a range of soil and aquatic habitats. *Flavobacterium* have previously been isolated from both red and brown marine algae as well as microalgae and numerous species have been shown to have the potential for degradation of a range of algal polysaccharides (Mann *et al*., 2013; Nedashkovskaya *et al*., 2014) . Additionally, the presence of *Flavobacterium* has been shown to increase flocculation in microalgae, aiding in the harvest of microalgae for biodiesel production (Lee *et al*., 2013; Lee *et al*., 2008). The addition of *Flavobacterium* (along with *Rhizobium*, *Hyphomonas*, *Terrimonas*, and *Mesorhizobium*) to a culture of the green algae *Chlorella vulgaris* was linked to an increase in biomass, with the suggestion that these bacteria were in a mutualistic relationship, fixing organic carbon released by microalgae whilst supplying inorganic and low molecular weight organic carbon, influencing algal growth and metabolism (Cho *et*

al., 2015). An earlier study, by Chirac *et al*. (1985) added *Flavobacterium* to axenic cultures of *B. braunii* which then produced higher levels of hydrocarbons and had increased biomass when compared to the axenic strain. Although we cannot determine which species is present in consortium with *B. braunii*, it looks likely that this is a beneficial member of the bacterial community.

Sample B had a higher proportion of *Flavobacterium* than sample A. *Flavobacterium* isolates are frequently found in diverse biofilm structures (Basson *et al*., 2008), and their presence here would indicate they are part of the bacterial community forming a biofilm on the surface of *B. braunii*.

Planctomycetes

The phyla Planctomycetes, Verrucomicrobia and Chlamydiae make up the PVC superphylum. The Planctomycetes phylum comprises of Gram negative budding bacteria which possess the unusual features of intracellular compartmentalisation and a lack of peptidoglycan in their cell walls (Fuerst, 2005). Planctomycetes are found in a wide range of aquatic and terrestrial environments, however the majority remain uncultivated and largely uncharacterised (Yoon *et al*., 2014; Fuerst & Sagulenko, 2011). Planctomycetes are often associated with macroalga where they have been found to form biofilms on the algal surface (Bengtsson and Øvreås, 2010; Lage and Bondoso, 2011). Factors that contribute to the successful colonisation of macroalgal surfaces by Planctomycetes have been identified: many species posses a holdfast, allowing for attachment to the algae; macroalgae also secrete a variety of sulfated polysaccharides which are used as a substrate by Planctomycetes in the production of sulfatases. Additionally, the unusual peptidoglycan-free cell wall of Planctomycetes enables them to remain unaffected by several different anti-microbial compounds released by macroalgae (Lage and Bondoso, 2014). There is little research available into the interactions between Planctomycetes and microalgae, such as *B. braunii*. However, as the cell walls of the green algae are rich in sulfated polysaccharides, the Plantomycetes present are likely to be involved in a symbiotic relationship with *B. braunii*.

Planctomycetes were found in both samples, but the highest proportion was in sample B (12.7 %) compared to sample A (3.1%), adding further weight to the argument that they are most likely forming a biofilm. In sample B QIIME further

classified the majority of reads from the Planctomycetes phylum to the order Phycisphaerales (10.5 %). The order Phycisphaerales contains three species of aquatic bacteria: *Algisphaera agarilytica, Phycisphaera mikurensis* and *Tepidisphaera mucosa.* It is difficult to determine which of these species are present in the *B. braunii* consortium; 16S rDNA sequences from the three species were looked for in the *B. braunii* consortium (using BLAST) and appeared to be present in equal quantities, it could therefore be said that all species are likely to be present. *A. agarilytica* and *P. mikurensis* have both previously been isolated from the marine algae *Cladophora* sp. and *Porphyra* sp. respectively (Yoon *et al*., 2013; Fukunaga et al., 2009) and *T. mucosa* has been isolated from hot springs in Russia (Kovaleva *et al.* 2015).

Proteobacteria

The most abundant phylum in both sample A and sample B is Proteobacteria (61.8% of sample A and 46.6% of sample B). Proteobacteria in sample A are composed of 11.3 % Alphaproteobacteria, 3.1% Betaproteobacteria, and 44.1 % Gammaproteobacteria (3.4% not classified beyond the phylum level). Proteobacteria in sample B are composed of 7.3% Alphaproteobacteria, 18.0% Betaproteobacteria, and 19.4% Gammaproteobacteria (1.8% not classified beyond the phylum level).

When looking further at the Gammaproteobacteria, 20.7 % of the reads in sample A are not classified beyond the class level, 8.9 % are classified into the order Legionellales and 13.8 % to the family Coxiellaceae (also in the order Legionellales). The Coxiellaceae contains three genera: *Aquicella, Coxiella* and *Diplorickettsia,* all of which are intracellular bacteria which form endosymbiotic relationships with other eukaryotes, such as ticks and protozoa (Mediannikov *et al*., 2010; Santos *et al*.,2003; Taylor et al., 2012). As with sample A, Coxiellaceae are found in similar numbers in sample B, with 10.5 % of reads assigned to this family as well as 3.8% assigned to the order Legionellales (remaining Gammaproteobacteria were not assigned beyond the class level). The need for Coxiellaceae to live within a host suggests that they are forming an endosymbiotic relationship with *B. braunii*, although they have not previously been documented living alongside algae.

The Alphaproteobacteria in both sample A and sample B is dominated by the order Rhizobiales (8.5 % of sample A and 3.8 % of sample B). A large range of taxa from the order Rhizobiales appear to be present in low numbers in both samples, with sample A having 1.4 % assigned to the genera *Hyphomicrobium* and *Devosia,* 1.5 % assigned to the family Bradyrhizobiaceae and the remaining reads assigned in numbers less than 1 % to the genera *Chelatococcus*, *Afipia*, *Bradyrhizobium*, *Filomicrobium*, *Pedomicrobium*, *Methylobacterium* and *Mesorhizobium.* Sample B has the same taxa present but none with more than 0.9 % of reads assigned to them. Three of the bacteria isolated from *B. braunii* in chapter 3 are from the order Rhizobiales (*Agrobacterium* sp. SUL3, *Shinella* sp. GWS1 and *Shinella* sp. SUS2). All three are from the family Rhizobiaceae. No reads have been assigned to the Rhizobiaceae for either sample, however there are numerous members of the order Rhizobiales present and it is possible that *Shinella* and *Agrobacterium* have been wrongly classified as one of those mentioned above (this is especially likely for *Shinella* due to a relatively low number of *Shinella* sequences in databases) or they may be in the small proportion of reads (0.8 % sample A, 0.6 % sample B) that have not been classified beyond the order level. These numbers would suggest that if present, *Shinella* and *Agrobacterium* are more likely to be in low numbers and in the planktonic bacterial population.

There is a more noticeable difference between the numbers of Betaproteobacteria in each sample, with 3.1 % and 18.% assigned to sample A and sample B respectively. The most abundant order present in both samples from the Betaproteobacteria is Burkholderiales, with 1.6 % and 12.1 % assigned to sample A and sample B respectively, and the majority of reads from this order have been further assigned to the family Comamonadaceae (1.5 % and 11.7 % of sample A and sample B respectively). Five genera from the family Comamonadaceae are present in sample B: *Methylibium*, *Aquincola, Pelomonas, Leptothrix* and *Mitsuaria* in numbers ranging from 1.9 - 2.3 %. The type species of the genus *Methylibium* is *Methylibium petroleiphilum,* a degrader of petroleum groundwater pollutants, the whole genome sequence of this bacterium also contains genes linked to aromatic hydrocarbon and alkane degradation (Kane *et al.,* 2007). The sole species of the genus *Aquincola* is *Aquincola tertiaricarbonis* which has been

previously isolated from groundwater contaminated with fuel oxygenates, where it was found to degrade methyl *tert*-butyl ether (MTBE) (Muller *et al.,* 2008). MTBE compounds are resistant to microbial degradation with only a few organisms capable of this, including *Aquincola tertiaricarbonis* and the previously discussed *Methylibium petroleiphilum* (Schäfer *et al.,* 2007).

Verrucomicrobia

Members of the phylum Verrucomicrobia are present in both sample A (loose association) and sample B (close association), with a larger proportion in sample B. The majority of reads assigned to this phylum have been further assigned to the genus *Prosthecobacter*, with 2.1 % of sample A and 13.2 % of sample B assigned to this genus. *Prosthecobacter* are freshwater Gram-negative, obligate aerobic bacteria. At the time of writing there are six known species of *Prosthecobacter*: *P. fusiformis* (Staley et al., 1976)*, P. dejongeii, P. debontii, P. vanneervenii* (Hedlund *et al.,* 1997)*, P. fluviatilis* (Takeda *et al*., 2008) *and P. algae* (Lee *et al.,* 2014). The *Prosthecobacter* genome has the unique feature of possessing tubulin-like genes; tubulins are thought to be unique to eukaryotes with the homolog *FtsZ* found in bacteria, however, with the exception of *P. fluviatilis*, all *Prosthecobacter* possess the genes encoding *BtubA* and *BtubB* which have higher sequence similarity to eukaryotic tubulin than to *FtsZ* (Schlieper et al., 2005). It has been hypothesised that the presence of *BtubA* and *BtubB* in Prosthecobacter is the result of horizontal gene transfer (Schlieper et al., 2005).

The use of 16S rDNA for taxonomic classification limits this study to determining bacteria to the genus level, therefore it is not possible to state which *Prosthecobacter* species are present in consortium with *B. braunii.* However, one species of interest to this study is *P. algae*, a species isolated from activated sludge using algal metabolites (Lee *et al*., 2014). Compounds found in algal metabolites, such as peptides, sterols and fatty acids can be used as carbon and nutrient sources for bacteria. *P. algae* was discovered in 2014 when Lee *et al*. took algal metabolites from a microalga (*Ankistrodesmus gracilis*) and used them in agar media to isolate microorganisms that could utilise algal metabolites. *Prosthecobacter* are present in highest numbers in sample B, bacteria in close

association with *B. braunii*, and it is possible that some of the bacteria in this group are also utilising algal metabolites.

3.6 Summary

The microbial population found in both loose and close association with *B. braunii* has been analysed through the use of 16S rRNA gene sequencing. Both parts of the microbial consortium were shown to have a diverse range of organisms and differences were observed between the community living in close association with
B. braunii and the community living in loose association. Bacteria were found living with *B. braunii* which have previously been associated with microalgae, including *Flavobacterium* and members of the phylum Plantomycetes.

Differences can be seen when using different classification tools to classify 16S rDNA sequence data. This chapter looked at three tools: Kraken, Megan, One Codex and QIIME which were used to classify a known microbial community. Based on rates of false negatives and false positives QIIME performed best, followed by Kraken. However, the main differences in classifications carried out using these tools was in the abundances of organisms classified, demonstrating that results from 16S rDNA sequence classifiers should be approached with caution when discussing relative abundances. Additionally, difference in taxonomic classification were seen when looking at different variable regions of the 16S rRNA gene. Three variable regions of the 16S rRNA gene were targeted for PCR amplification and sequencing: V1-V2, V3-V4 and V5-V8; the V5-V8 region was poor at classifying organisms to a low taxonomic level, whilst the V1- V2 and V3-V4 regions were able to classify larger proportions of reads to the genus level.

Information regarding taxonomic classifiers and variable regions will be used when planning analysis in the next chapter, which will look at a 16S rDNA sequence dataset taken from a very different environment – acid mine drainage.

Chapter four: Assessing the complexity of the microbial community found in acid mine drainage from Wheal Jane and Wheal Maid using 16S rRNA gene sequencing

4.1 Introduction

4.1.1 An introduction to Wheal Jane and Wheal Maid, Cornwall

Cornwall has a long rich history of metalliferous mining dating back to the Bronze Age. At its peak in the 1800s Cornwall was the world's largest producer of tin and copper (Camm *et al*., 2004; Pirrie *et al*., 2003) while smaller quantities of lead, zinc, tungsten, bismuth, antimony, cobalt, uranium, nickel, manganese, iron, silver and arsenic were also recovered from Cornwall's mines (Camm *et al*., 2003). The collapse of the world tin cartel in 1985 was the beginning of the end for Cornish mining and the last working mine, the South Crofty tin mine, closed in 1998 (Adams & Younger, 2002). However, the impact of mining on the Cornish environment remains evident, with the Environment Agency calling abandoned mines "one of the most significant pollution threats in Britain" (Johnston *et al*., 2008). The main pollutant from abandoned mines in the Cornwall area is acid mine drainage (AMD) As discussed in Chapter 1, AMD contaminates areas with various pollutants, including lead, arsenic, mercury and other metals/metalloids. The threat that AMD poses to the environment was demonstrated with devastating consequences in 1992 when, a year after its closure, the Wheal Jane mine (Grid reference: SW772426) released 45 million litres of AMD into the local water course causing severe contamination of the River Carnon, the Fal River and the Fal estuary (Neal *et al*., 2004). The incident was called an environmental disaster and resulted in a full scale mine water treatment plant being built on the site which remains there today, pumping AMD out of the mine at a rate of 110 litres a second (Younger *et al*., 20014).

At the active treatment plant, once pumped from the mine AMD is dosed with lime slurry and aerated, raising the pH and causing metals to precipitate out of solution into the lime sludge which is then removed and disposed of while the supernatant is discharged into local rivers (Brown *et al*., 2007). Alongside this active treatment of AMD at the Wheal Jane site, investment has been made in passive treatment options to investigate the long-term treatment of AMD at both this and other sites across the UK. The Wheal Jane passive treatment plant consists of aerobic reed beds designed to remove iron and arsenic, an

anaerobic cell to encourage the reduction of sulphate and the removal of zinc, copper, cadmium and the remaining iron, and aerobic rock filters designed to promote the growth of algae and facilitate the precipitation of manganese (Whitehead *et al*., 2004). The Wheal Jane passive treatment plant has led the way for alternative, more sustainable treatment options and this remains an active area of research.

Three km from Wheal Jane is the site of Wheal Maid (Grid reference: SW749423). The Wheal Maid site is approximately 0.08 km^2 and consists of two tailings lagoons separated by three dams. The site was operated during the 1970s and 1980s and the lagoons contain approximately 220 000 m^3 of fine grained mineral processing waste (tailings) from nearby mines (Veen *et al*., 2016). In 2008 Carrick District Council produced a Schedule of Determination declaring the Wheal Maid site as contaminated land that posed a significant threat to the health of children who frequently use mountain bike tracks that run around the lagoons as well as posing a risk to controlled waters through the leaching of metals from the site. The pollutants identified in AMD at Wheal Maid included arsenic, cadmium, chromium, copper, iron, lead, nickel and zinc (Carrick District Council, 2008). The site is currently monitored by the Environment Agency but there is no active treatment of the area.

As discussed in chapter one, microbial communities are known to have a significant effect on the production and management of AMD. Studies into the microbial population at Wheal Jane have mainly focussed on the wetland ecosystem that forms part of the passive treatment plant (Barley *et al*., 2005; Hallberg & Johnson, 2005; Whitehead & Prior, 2005), and to date there have not been any studies that have profiled the microbial community present in Wheal Maid. An understanding of the microbial ecology of this area is important if all bioremediation options are to be fully explored. The two sites provide different environments from which to study AMD microorganisms. Wheal Jane has a complex structure; the modern workings within the mine extend to around 450 m and the mine is surrounded by and interconnected with older disused mines. Water flowing through Wheal Jane is likely to enter from these connected workings with the exact source remaining unknown. Wash-off from a variety of environments in the surrounding area could therefore be entering the

Wheal Jane mine creating an environment that is likely to change depending on rainfall levels. Wheal Maid is a stagnant pool, creating an environment that is less changeable than Wheal Jane, although it is still affected by local weather conditions.

4.1.2 Aims

The aim of this chapter is to carry out a preliminary molecular microbiology study on the Wheal Jane and Wheal Maid site, using 16S rRNA gene sequencing to gain a better understanding of the complexity of the microbial communities present and identify any keystone species which may be of further interest, especially when looking at the bioremedial potential of the microbial community. This initial study looking at the complexity of the microbial communities at Wheal Maid and Wheal Jane will allow for the future planning of further studies, such as full shotgun metagenomics of the site.

4.2 Materials and methods

4.2.1 DNA extraction and sequencing

AMD water samples from Wheal Maid and Wheal Jane were obtained by Holly Smith-Baedorf from Plymouth Marine Laboratories. Smith-Baedorf carried out DNA extractions from water samples using the Sigma-Aldrich Bacterial Genomic Miniprep kit. Sediment samples from Wheal Maid were obtained by Chris Bryan from The University of Exeter (Environment and sustainability institute, Penryn Campus). Samples were taken from two locations at Wheal Maid (Figure 4.1), and from three depths (Depth 1 = Surface, Depth 2 = 30 cm, Depth 3 = 50 cm). Bryan carried out DNA extractions from sediment samples using the Mo Bio PowerSoil DNA isolation kit. PCR reaction mix used the NEBnext High-Fidelity PCR master mix and primers designed for the V3-V4 region of the 16S rRNA gene. A second PCR phase was incorporated to add flowcell binding regions, an Illumina adapter and a multiplexing barcode, creating DNA libraries which were combined and sequenced using the Illumina MiSeq. 300 bp paired-end sequence reads were obtained and demultiplexed by the Exeter Sequencing Service. Paired end reads were trimmed using Trim Galore prior to analysis.

Figure 4.1 The two sites at Wheal Maid from which sediment samples were obtained. Site one = Oxidised sediment below the water level, Site 2 = Grey material rarely below the water level. (Photo credit: Chris Bryan, University of Exeter)

4.2.2 Bioinformatics tools and software

Table 4.1 shows software and websites used in this study. Default parameters were used for all programs unless stated otherwise below.

4.2.3 Taxonomic classification and phylogeny

Kraken was used for initial analysis of 16S rRNA gene sequence data from Wheal Maid and Wheal Jane using the paired reads in FastQ format. The standard Kraken database was used. Pavian was used via R studio to visualise and analyse taxonomic information.

Paired end reads from Wheal Maid sediment samples were joined using FLASH and QIIME was used for taxonomic classification.

QIIME was run using the following workflow:

1. add_qiime_labels.py
2. pick_otus.py
3. pick_rep_set.py
4. assign_taxonomy.py
5. make_otu_table.py
6. filter_samples_from_otu_table.py –n 2
7. normalize_table.py
8. summarize_taxa_through_plots.py

BLAST alignments were carried out on the NCBI server using both the 16S rRNA Bacteria and Archaea database and the nucleotide database with default parameters and the blastn program.

The SeaView package was used for phylogenies as follows: Sequences were aligned using MUSCLE, with poorly aligned and divergent regions eliminated using Gblocks. Maximum likelihood phylogenetic trees were constructed using PhyML, with the GTR substitution model and bootstrapping of 100.

Table 4.1. Software and websites used in this study

Name	**Version**	**Available from:**	**Reference**
NCBI BLAST (online server)		https://blast.ncbi.nlm.nih.gov/	
FLASH	1.2.7	https://ccb.jhu.edu/software/FLASH/	Magoc & Salzberg, 2011
Kraken	0.10.6	https://ccb.jhu.edu/software/kraken/	Wood & Salzberg, 2014
Pavian	0.6.2	https://github.com/fbreitwieser/pavian	Breitwieser & Salzberg,
QIIME (Virtual Box)	1	http://qiime.org/install/virtual_box.html	Caporaso *et al*., 2010
SEAVIEW	4.5.3	http://doua.prabi.fr/software/seaview	Gouy *et al*., 2010
Trim galore	0.3.3	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/	Krueger, 2015

4.3 An initial comparison indicates a less complex bacterial community in Wheal Maid than that of Wheal Jane

To carry out an assessment of the complexity of the microbial populations living in acid mine drainage in the Wheal Jane mine and the Wheal Maid tailings lagoon, samples of water were taken from each site, DNA extractions carried out and 16S rRNA gene sequences obtained. These samples were taken over the course of four months to try and determine how the microbial populations change with time. However due to problems with the quantity and quality of DNA extracted from the samples, 16S rRNA gene sequence data was only obtained from three samples from each location, sampled on the following dates:

Wheal Maid

3/3/15 (referred to from here on as Wheal Maid sample 1)

31/3/15 (referred to from here on as Wheal Maid sample 2)

21/4/15 (referred to from here on as Wheal Maid sample 3)

Wheal Jane

24/2/15 (referred to from here on as Wheal Jane sample 1)

31/3/15 (referred to from here on as Wheal Jane sample 2)

21/4/15 (referred to from here on as Wheal Jane sample 3)

Therefore the time period covered is 7 weeks for Wheal Maid and 8 weeks for Wheal Jane. Number of reads obtained for each sample are shown in Table 4.2.

4.3.1 Wheal Maid and Wheal Jane have differences in the complexity of their communities

Taxonomic classification for each sample was carried out using Kraken and Pavian was used for visualisation of results. Kraken was chosen as it has been shown to perform well for 16S rRNA gene sequence classification and is relatively quick to run (Chapter three, 3.3), making it a good choice for an initial analysis of the complexity of the two sites. Figures 4.2-4.7 show the results of taxonomic classification.

Table 4.2. Number of reads and percent classified using Kraken for Wheal Jane and Wheal Maid water samples.

Sample	Number of raw reads	Percentage of reads classified (Kraken)
Wheal Jane 1	856232	88.2 %
Wheal Jane 2	1185938	88.1 %
Wheal Jane 3	1296548	89.6 %
Wheal Maid 1	925687	94.2 %
Wheal Maid 2	907511	95.2 %
Wheal Maid 3	1169788	90.6 %

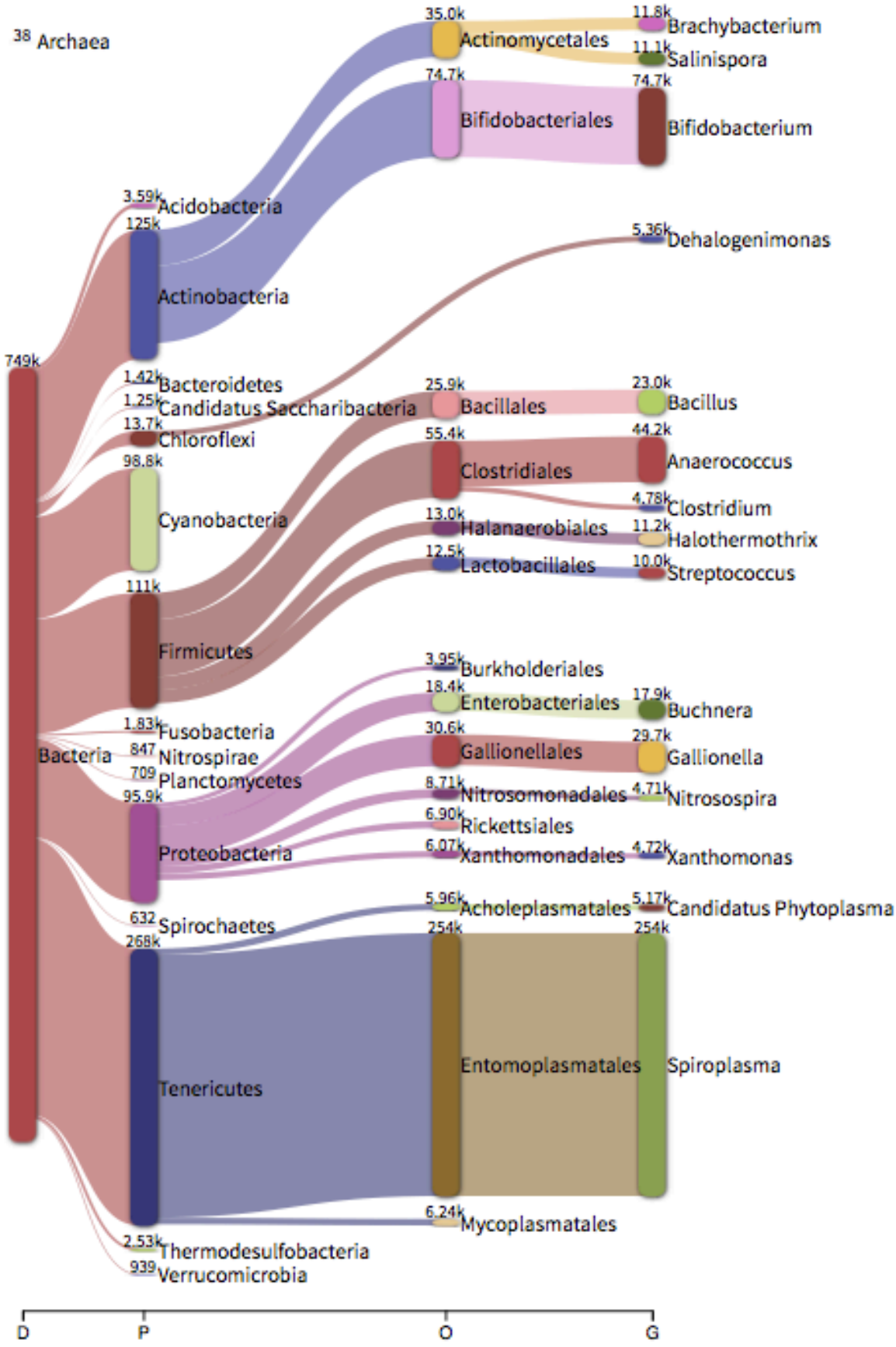


Figure 4.2, taxonomic distribution for Wheal Jane, sample one. D= Domain, P=Phylum, O=Order and G = genus

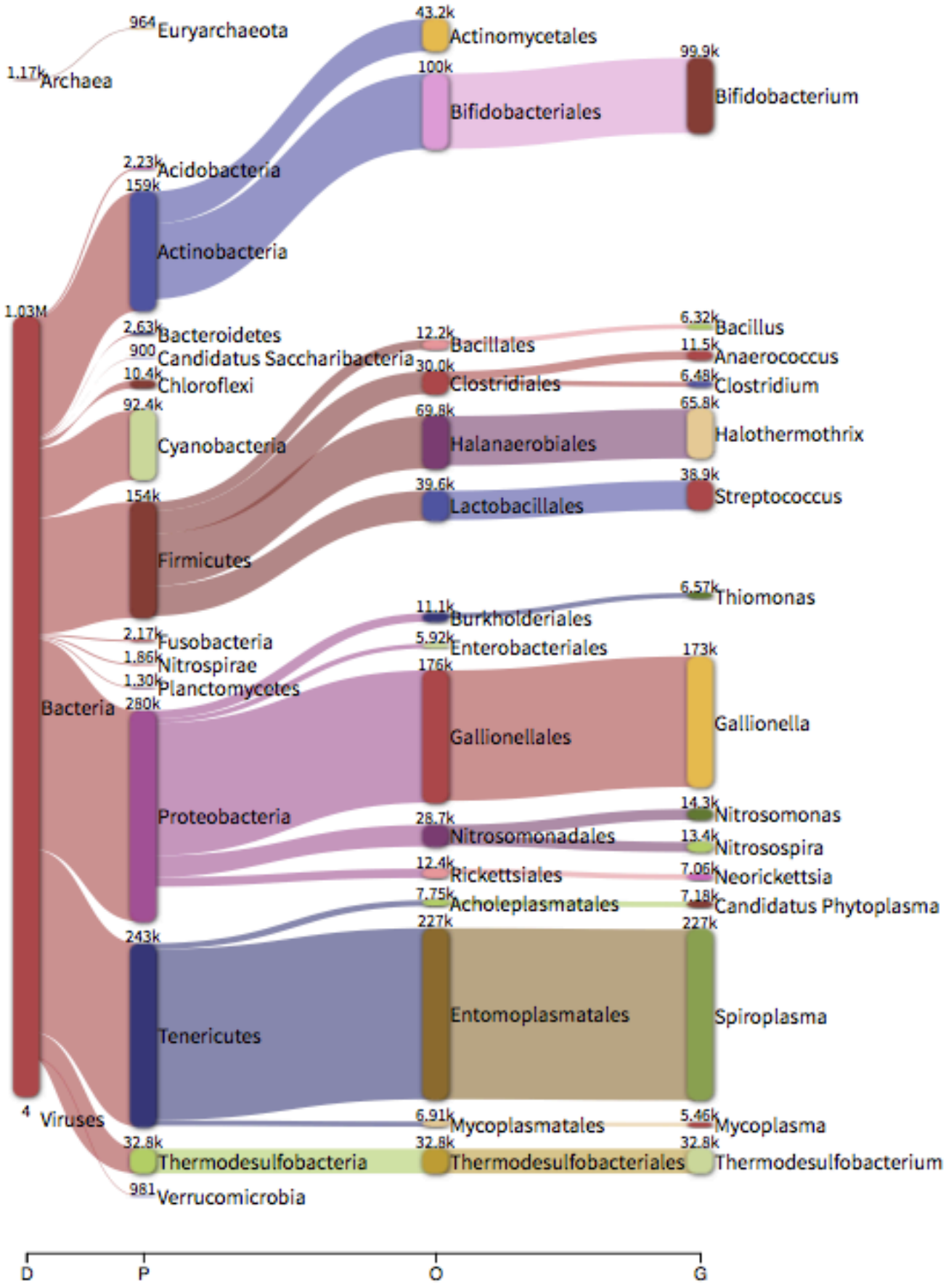


Figure 4.3, taxonomic distribution for Wheal Jane, sample two. D= Domain, P=Phylum, O=Order and G = genus

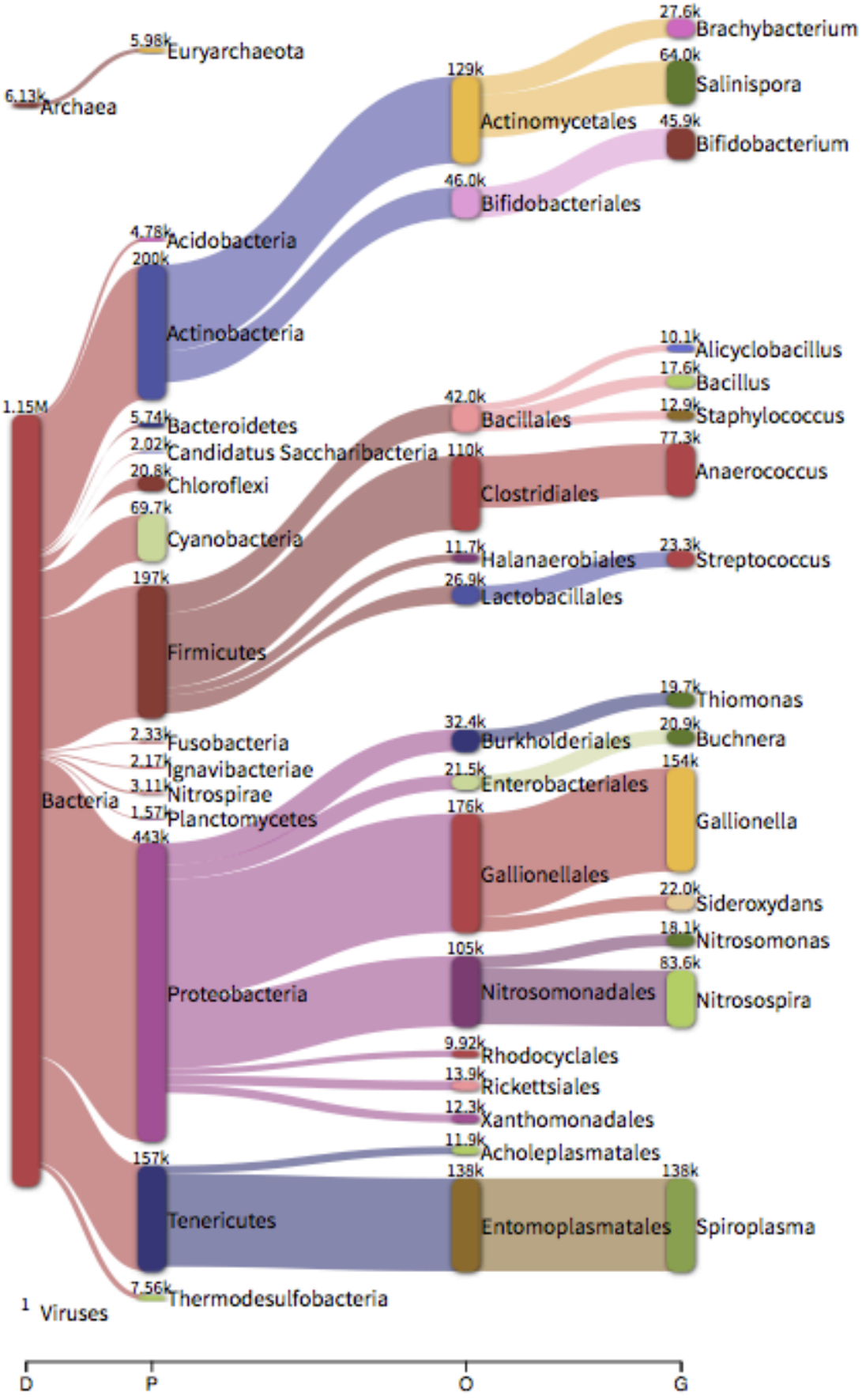


Figure 4.4, taxonomic distribution for Wheal Jane, sample three. D= Domain, P=Phylum, O=Order and G = genus

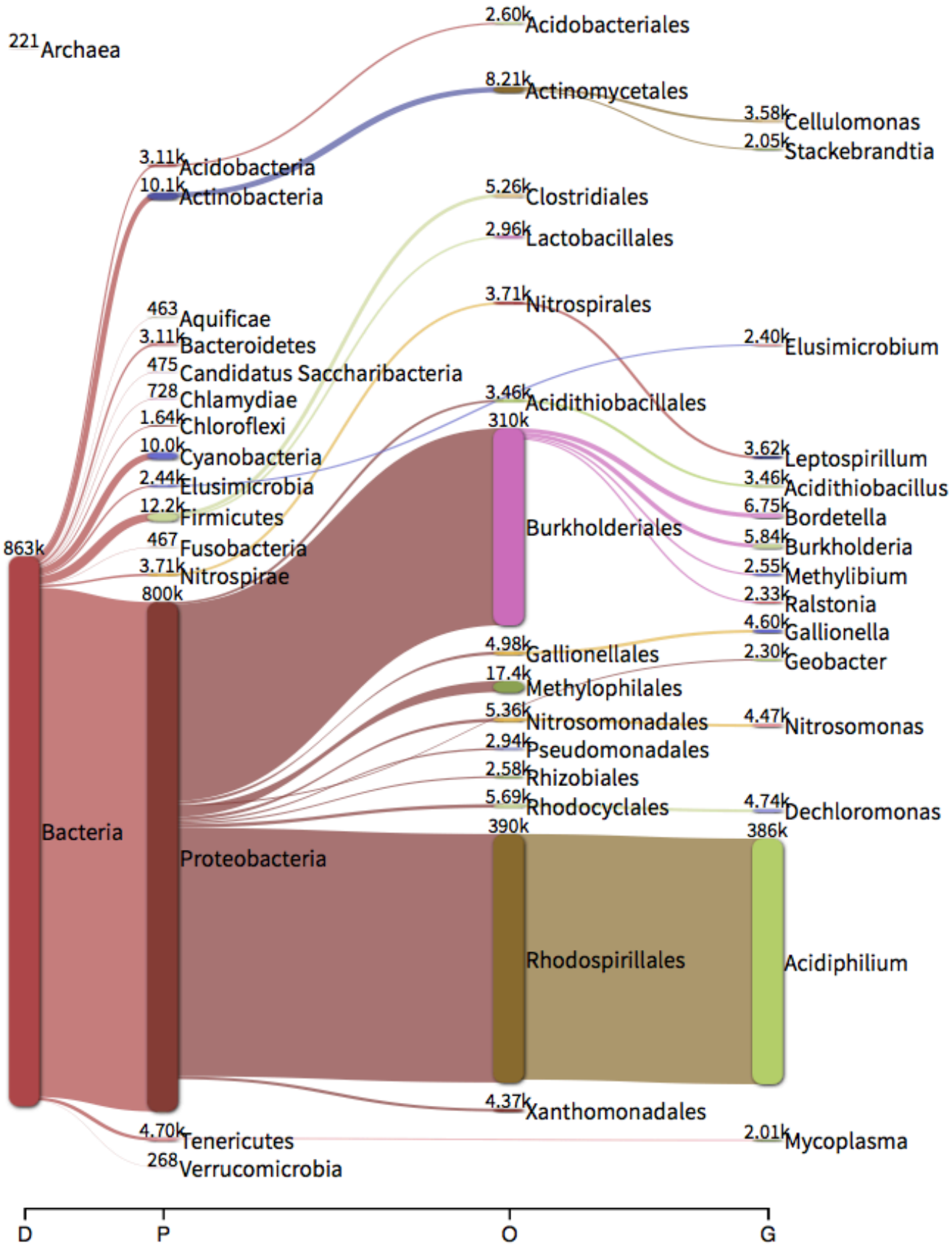


Figure 4.5, taxonomic distribution for Wheal Maid, sample one. D= Domain, P=Phylum, O=Order and G = genus

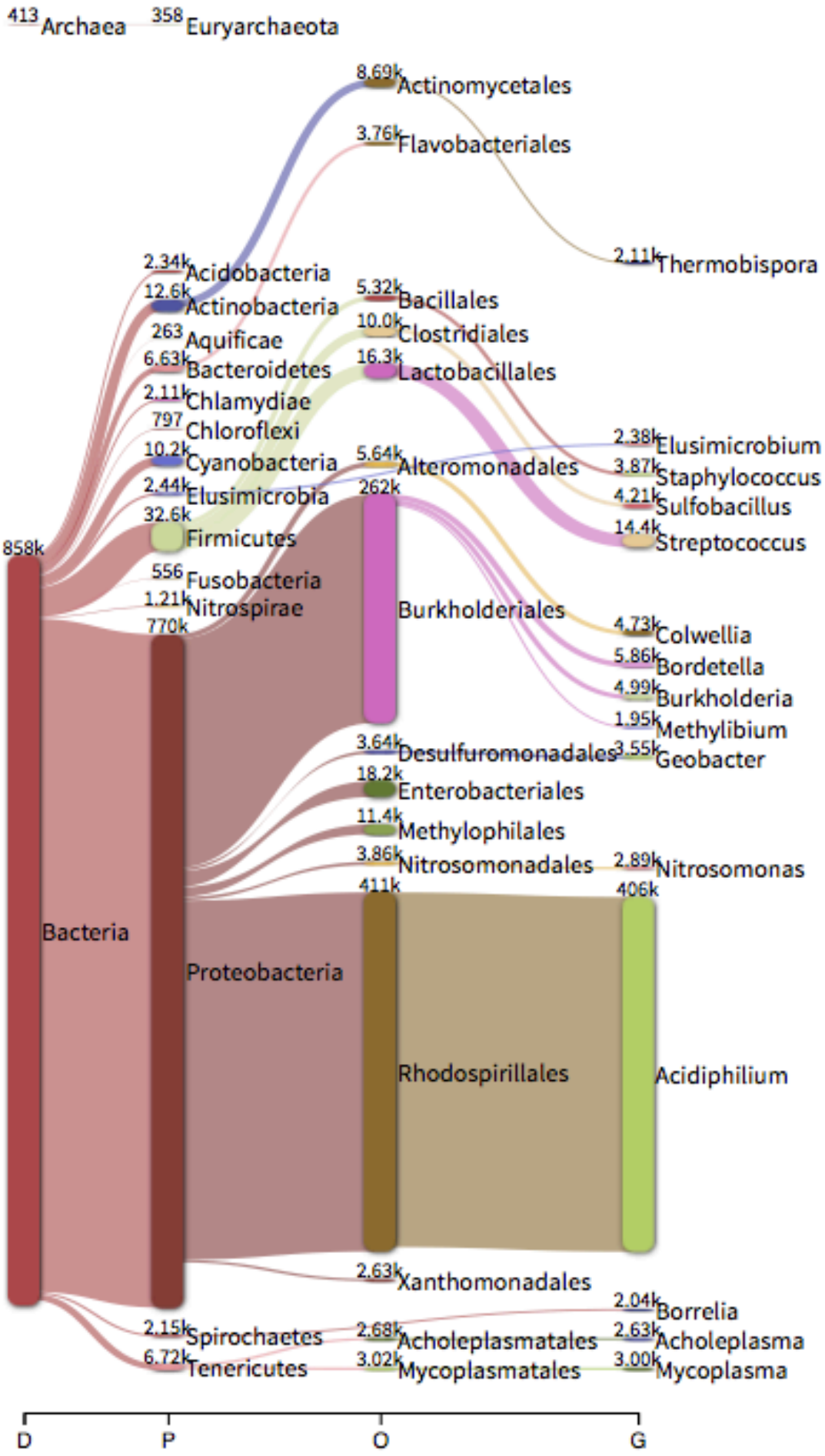


Figure 4.6, taxonomic distribution for Wheal Maid, sample two. D= Domain, P=Phylum, O=Order and G = genus

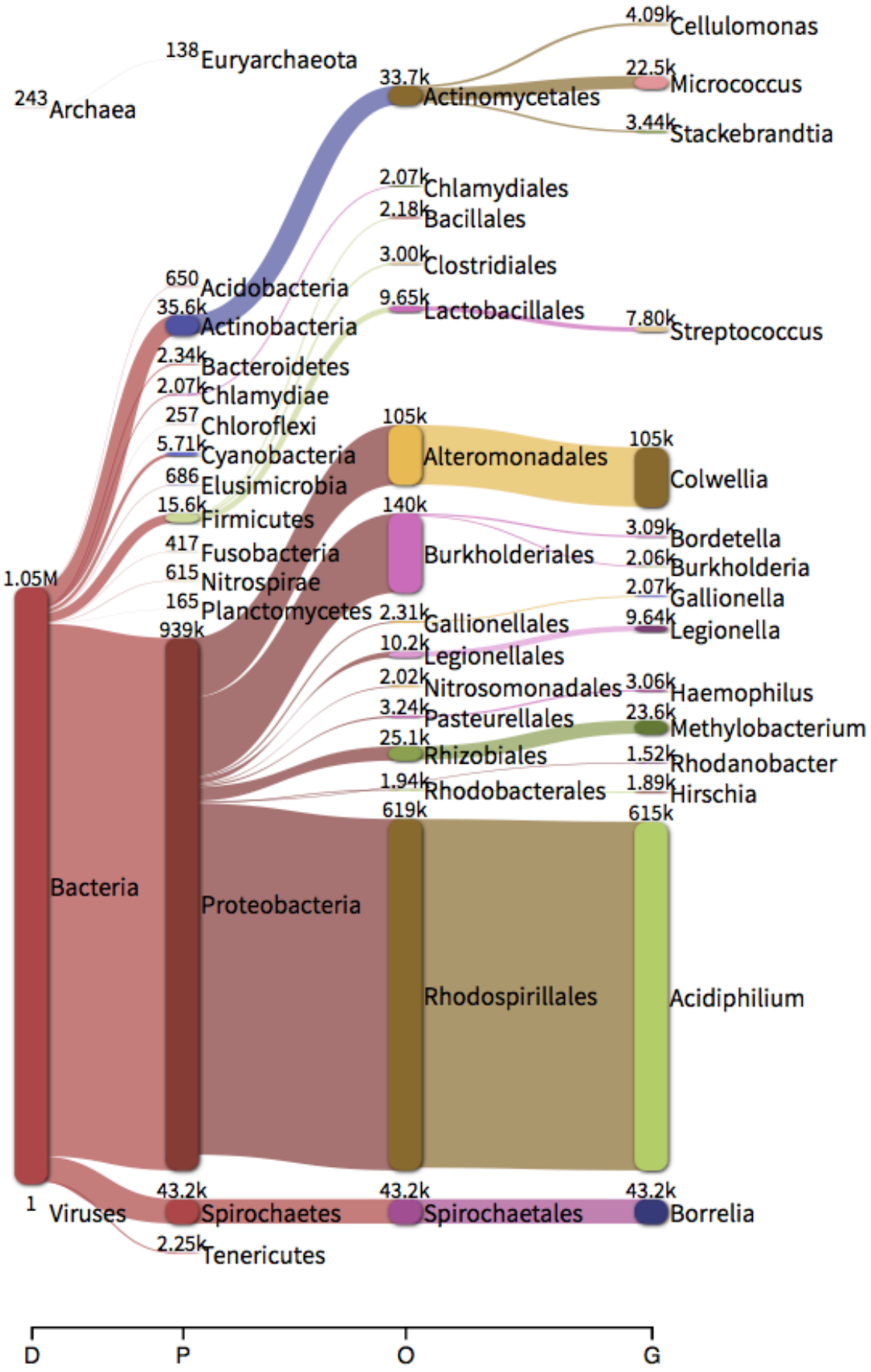


Figure 4.7, taxonomic distribution for Wheal Maid, sample three. D= Domain, P=Phylum, O=Order and G = genus

Between 5-10 % of reads from Wheal Maid samples were unclassified and between 10-12 % of reads for Wheal Jane were unclassified. The 15 most abundant taxa at each taxonomic level below domain are displayed for each sample. There is a clear difference in the taxonomic distribution between Wheal Maid and Wheal Jane. The order Rhodospirillales is the dominant order in all three of the Wheal Maid samples with 48 %, 51 % and 61 % of samples 1,2 and 3 assigned to this order respectively; 99 % of all reads assigned to this order have been further assigned to the genus *Acidiphilium.* As the name suggests, members of the genus *Acidiphilium* are often found in acidic environments and can tolerate a wide range of pH levels. Several strains of *Acidiphilium* have previously been isolated from a range of AMD sites with several species demonstrating a resistance to heavy metals including copper, nickel and zinc (Auld *et al*., 2013). The second most abundant order in Wheal Maid is Burkholderiales with 38 %, 33 % and 14 % of reads from samples 1, 2 and 3 assigned to this order respectively. 93 – 94 % of reads assigned to Burkholderiales are not assigned beyond the order level. In Wheal Maid sample 3 the proportion of Burkholderiales is only 14 % whilst 10 % of reads are assigned to Alteromodales (very low numbers of reads are assigned to this order in samples 1 and 2); these reads are further assigned to *Colwellia,* a genus which is composed of psychrophilic, piezophilic bacteria commonly found in the depths of the ocean and unlikely to be present in the Wheal Maid environment, making this likely to be a misclassification. To test this, random reads from the Wheal Maid sequencing dataset which had been assigned to Alteromodales were used in a BLAST alignment against the NCBI 16S rRNA database, where they were found to have greatest similarity with the cyanobacterium *Stanieria cyanosphaera,* and in a BLAST alignment against the NCBI nucleotide database where they were found to have greatest similarity with a number of uncultured environmental clones. A phylogenetic tree (Figure 4.7) shows that reads assigned to *Colwellia* are most closely related to uncultured bacteria CN7 and CN10 which were obtained from acidic pit lakes in Spain (Santofimia *et al*., 2013). This would indicate that novel acidophilic species are present and have been misclassified. Other acidophilic bacteria present in smaller proportions in the Wheal Maid samples include the orders Acidobacteriales and Acidithiobacillales. *Leptospirillum,* which is known to be

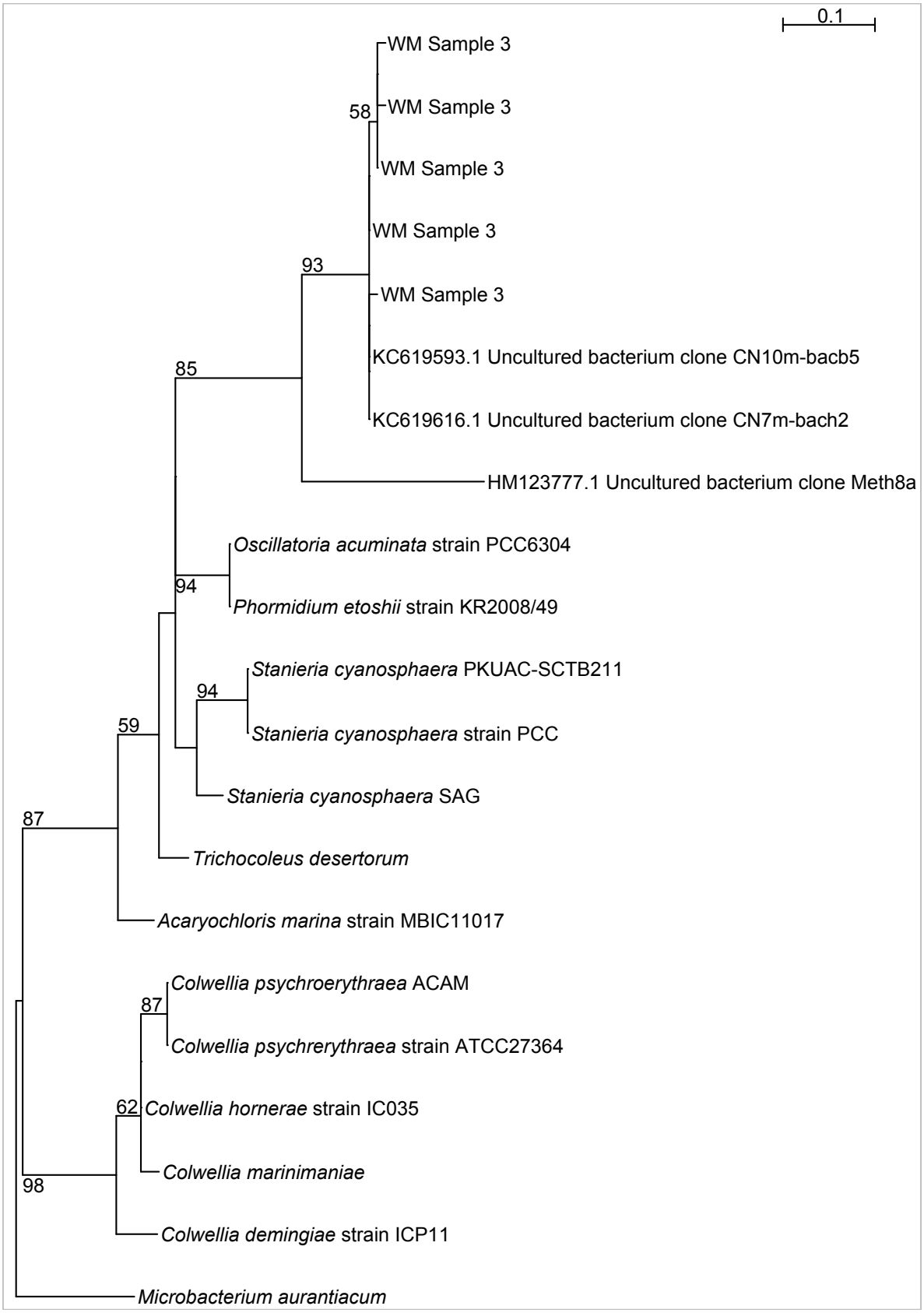


Figure 4.8, Maximum likelihood tree showing reads from wheal Maid sample three which had been assigned to the genus Colwellia, organisms classed as similar to these reads by BLAST and members of the genus Colwellia. The tree is rooted at *Microbacterium aurantiiacum* and bootstrap values over 50 are shown.

common in AMD, is present in all three samples but only at levels of around 1 %.

There is a high proportion of reads assigned to the genus *Spiroplasma* in the Wheal Jane samples, with 34 %, 22 % and 12 % of reads from samples 1, 2 and 3 assigned to this genus respectively. Members of the genus *Spiroplasma* are parasitic bacteria typically found in insect guts, plant phloem or human hosts and are not usually found in AMD (Herren *et al*., 2011), making their presence here in large quantities unlikely. After taking a number of random reads from the Wheal Jane sequencing dataset which had been assigned to this genus and using them in a BLAST alignment against the NCBI 16S rRNA and nucleotide databases they were shown to have a greater similarity to a number of uncultured bacteria. A phylogenetic tree (Figure 4.8) created using reads assigned to *Spiroplasma* along with organisms shown through BLAST to have high levels of similarity to these reads and members of the *Spiroplasma* genus shows they are more closely related to these uncultured bacteria than to *Spiroplasma*. Uncultured bacteria in Figure 5.8 were detected in samples obtained from uranium and heavy metal contaminated soil, acidic sulphuric wetlands and from benthic sediments. This indicates that, like Wheal Maid, samples from Wheal Jane are likely to contain novel extremophiles which have been misclassified. The genus *Gallionella* is present in Wheal Jane in samples 1, 2 and 3 at levels of 4 %, 17 % and 13 % respectively. Bacteria of the genus *Gallionella* are typically iron-oxidising acidophiles and have been previously isolated from AMD, where they thrive at pH >3 and at iron(II) concentration of >4 mM (Jones *et al*., 2015).

With the exceptions of those discussed above, Wheal Jane does not appear to contain high numbers of microorganisms typically found in AMD. As previously mentioned (section 5.1), the Wheal Jane mine contains water that has drained into the modern mine shaft workings from surrounding mines and water is flowing at a high rate. Water draining into Wheal Jane may have travelled through a variety of environments and it is likely to bring microbial life with it that does not naturally thrive in AMD. In order to fully understand the microbial population of Wheal Jane a 16S study using more samples, obtained over a longer time period, including numerous biological and technical replicates should be carried out.

Contamination of Wheal Maid water by bacteria which cannot thrive in AMD is also to be expected due the openness of the area which is likely to be frequented by wildlife and people. Additionally, run off from the surrounding area after heavy rainfall will occur. To reduce the amount of non-native organisms in samples from AMD at Wheal Maid samples could be obtained from the sediment rather than the water in the tailings pool. This will be looked at in the next section.

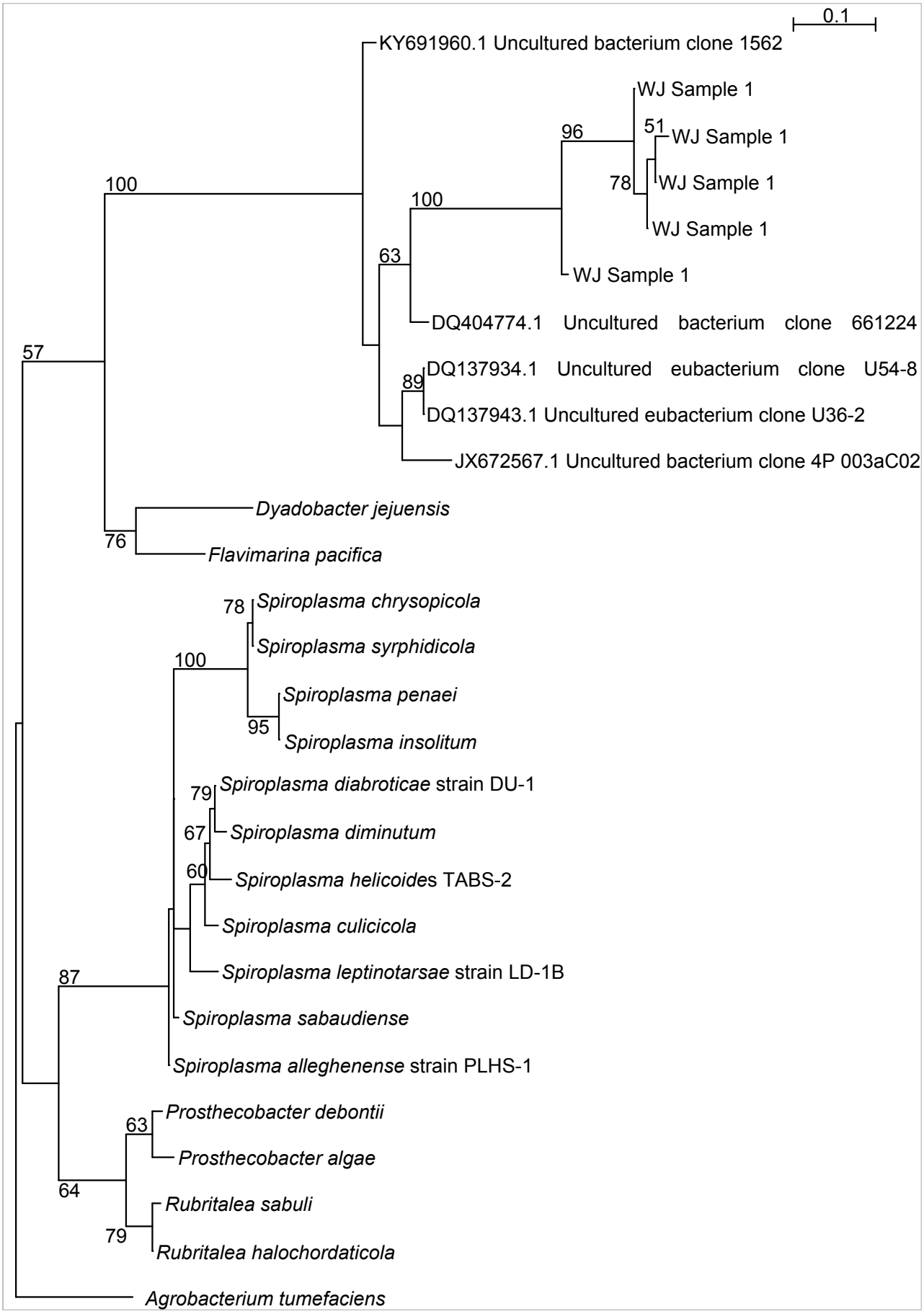


Figure 4.9, Maximum likelihood tree showing reads from Wheal Jane sample one which had been assigned to the genus *Spiroplasma*, organisms classed as similar to these reads by BLAST and members of the genus *Spiroplasma*. The tree is rooted at *Agrobacterium tumefaciens* and bootstrap values over 50 are shown.

4.4 Analysis of the bacterial population found in Wheal Maid sediment reveals a diversity of bacteria characteristic of global AMD sites.

As discussed in 4.3, the Wheal Maid tailings lagoon offers the chance to study microorganisms in a relatively stable AMD environment. By looking at sediment samples from different depths, organisms living in different micro-environments within the Wheal Maid lagoon can be compared. Samples were taken from two locations at Wheal Maid: Site 1, composed of oxidised sediment below the Winter water level and Site 2, composed of grey material rarely below water level (Figure 4.1). Samples were taken from the surface level (depth 1) and at depths of 30 cm (depth 2) and 50 cm (depth 3). Sequence statistics are shown in Table 4. 3. Moisture, pH and geochemical data was also obtained for the two sites (Tables 4.4 and 4.5). Although there is variation in conditions between the two sites and across the different depths, both sites contain metals which are typical of AMD, including: aluminium, iron, copper, zinc, arsenic, manganese, lead and cadmium as well as rare earth elements scandium, yttrium, gadolinium, terbium, dysprosium, holium, erbium, thulium, ytterbium and lutetium and radioactive elements thorium and uranium. The two sites also have similar pH ranges of 1.78 – 2.55 at site 1 and 1.62 – 2.96 at site 2. Moisture levels at the three depths are higher at site 1, which is unsurprising as this sits below the Winter water level, with a range of 20.9 - 25.7 %; site 2 has moisture levels at the three depths in the range of 7.1 - 22.2%.

5.4.1 Novel organisms may have been misclassified

QIIME was used for the analysis of all Wheal Maid sediment samples. QIIME was chosen as it was previously demonstrated to be a good choice for the analysis of 16S rRNA sequence datasets (Chapter three, 3.3). Figures 4.10-4.15 show output from QIIME. Reads were classified to the genus level where possible, although large numbers were only classified as far as class, order or family. Figures 4.10 and 4.11 show taxonomic differences between depths for the two sites. Figures 4.12 and 4.13 give distributions at the level of phylum, and Figures 4.15 and 4.16 give distributions at the lowest taxonomic level.

Table 4.3, Numbers of reads and percentage classified for Wheal Maid sites 1 and 2 at three depths. S1 = Site one, S2 = Site two,

Sample	Number of raw reads	Percentage of reads classified
WM S1, Depth 1	1412667	99.9 %
WM S1, Depth 2	5441703	99.9 %
WM S1, Depth 3	2663625	99.8 %
WM S2, Depth 1	5327845	99.9 %
WM S2, Depth 2	4638914	99.8 %
WM S2, Depth 3	3550629	99.9 %

Table 4.4 The chemical composition of sediment taken from two sites and three depths at Wheal Maid. Depth 1 = Surface, Depth 2 = 30cm, Depth 3 = 50cm.Data provided by Dr Chris Bryan, University of Exeter.

	Al (%)	Fe (%)	Cu (%)	Zn (%)	As (%)	Mn (ppm)	Ni (ppm)	Pb (ppm)	Cd (ppm)	Sc (ppm)
Site 1 Depth1	4.058	28.983	0.435	0.243	1.580	74.499	37.087	188.035	9.505	2.773
Site 1 Depth 2	3.485	6.066	0.016	0.007	0.249	204.437	8.024	75.019	0.127	3.632
Site 1 Depth 3	4.065	7.474	0.014	0.007	0.171	196.499	7.025	71.353	0.326	4.487
Site 2 Depth 1	4.643	7.338	0.010	0.007	0.156	208.553	4.403	106.841	0.209	5.685
Site 2 Depth 2	3.153	21.865	0.022	0.028	0.200	163.729	55.187	593.684	0.750	4.353
Site 2 Depth 3	3.620	3.078	0.014	0.012	0.075	71.413	6.267	191.444	0.247	4.280

	Y (ppm)	Gd (ppm)	Tb (ppm)	Dy (ppm)	Ho (ppm)	Er (ppm)	Tm (ppm)	Yb (ppm)	Lu (ppm)	Th (ppm)	U (ppm)
Site 1 Depth1	16.933	6.305	0.863	4.481	0.775	2.024	0.274	1.675	0.256	3.704	10.721
Site 1 Depth 2	3.483	0.980	0.132	0.742	0.140	0.413	0.062	0.427	0.069	2.206	1.385
Site 1 Depth 3	3.998	1.298	0.394	1.030	0.393	0.675	0.314	0.702	0.323	2.846	1.331
Site 2 Depth 1	5.562	1.409	0.305	1.211	0.336	0.768	0.218	0.832	0.232	2.732	1.285
Site 2 Depth 2	4.856	1.812	0.393	1.211	0.377	0.723	0.270	0.743	0.282	2.774	0.910
Site 2 Depth 3	5.738	1.558	0.204	1.081	0.210	0.595	0.091	1.284	1.284	1.284	1.284

Table 4.5 pH and moisture levels at two sites, three depths at Wheal Maid. . Depth 1 = Surface, Depth 2 = 30cm, Depth 3 = 50cm. Data provided by Dr Chris Bryan, University of Exeter

	pH	Moisture
Site 1 Depth 1	2.47	20.9%
Site 1 Depth 2	2.55	25.7%
Site 1 Depth 3	1.78	23.2%
Site 2 Depth 1	1.62	7.1%
Site 2 Depth 2	2.0	13.6%
Site 2 Depth 3	2.96	22.2%

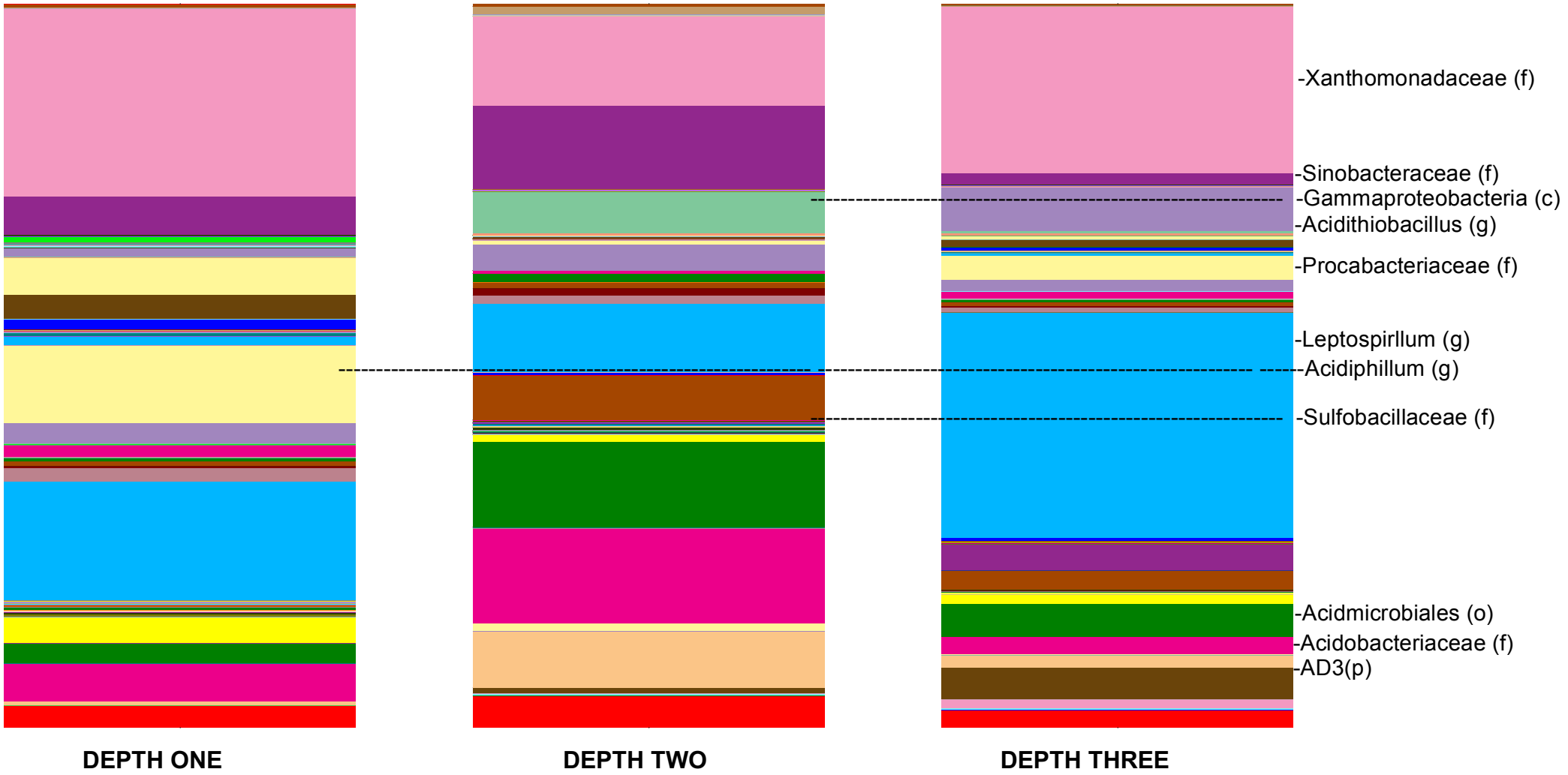


Figure 4.10 QIIME output demonstrating differences in taxonomic distribution across three depths at Wheal Maid site 1. Same coloured blocks across the samples represent the same taxa. Lowest possible taxonomic levels down to genus have been assigned. Taxa with at least 5 % assigned to them (at any depth) are labelled.

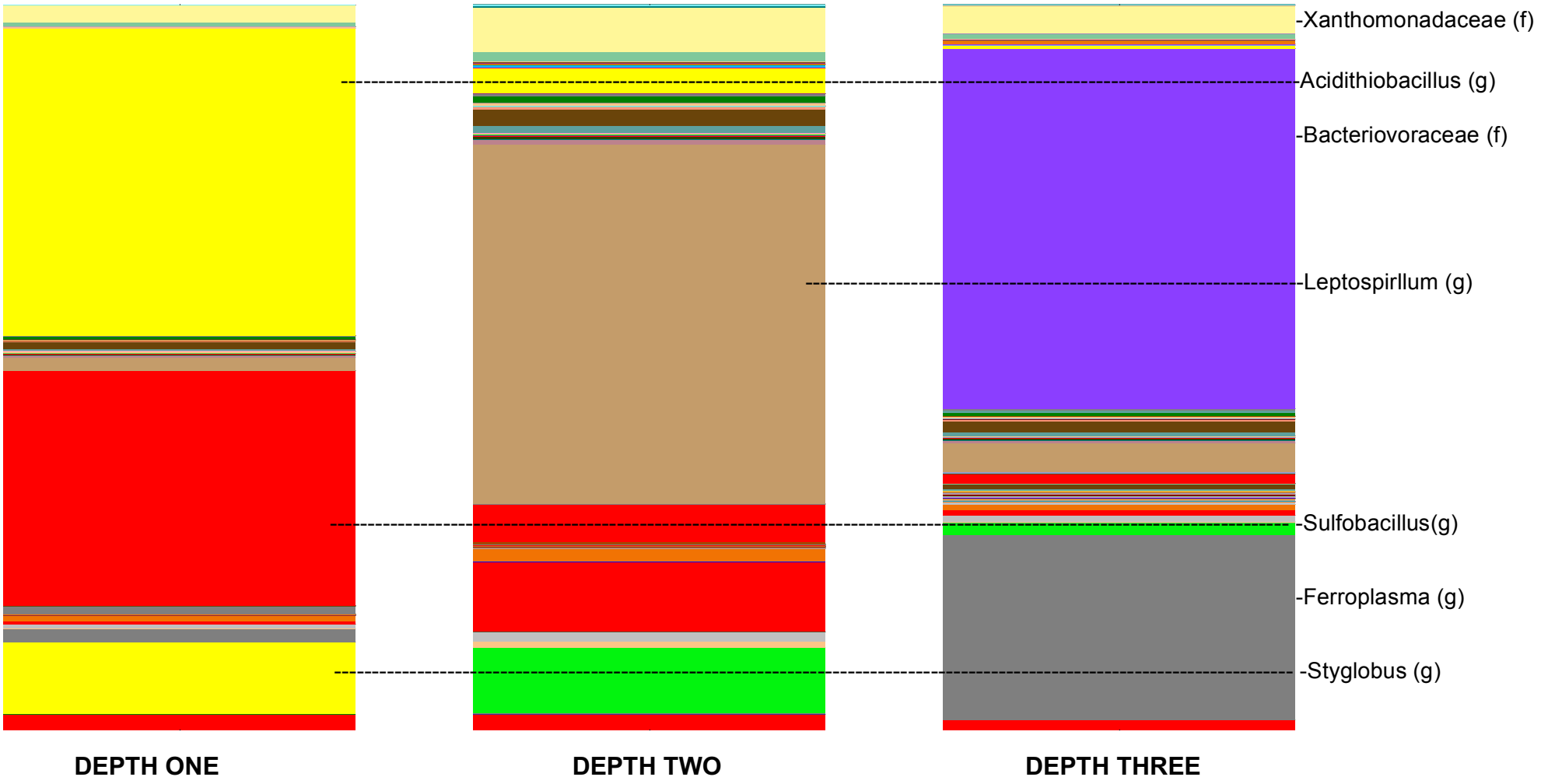


Figure 4.11 QIIME output demonstrating differences in taxonomic distribution across three depths at Wheal Maid site 2. Same coloured blocks across the samples represent the same taxa. Lowest possible taxonomic levels down to genus have been assigned. Taxa with at least 5 % assigned to them (at any depth) are labelled

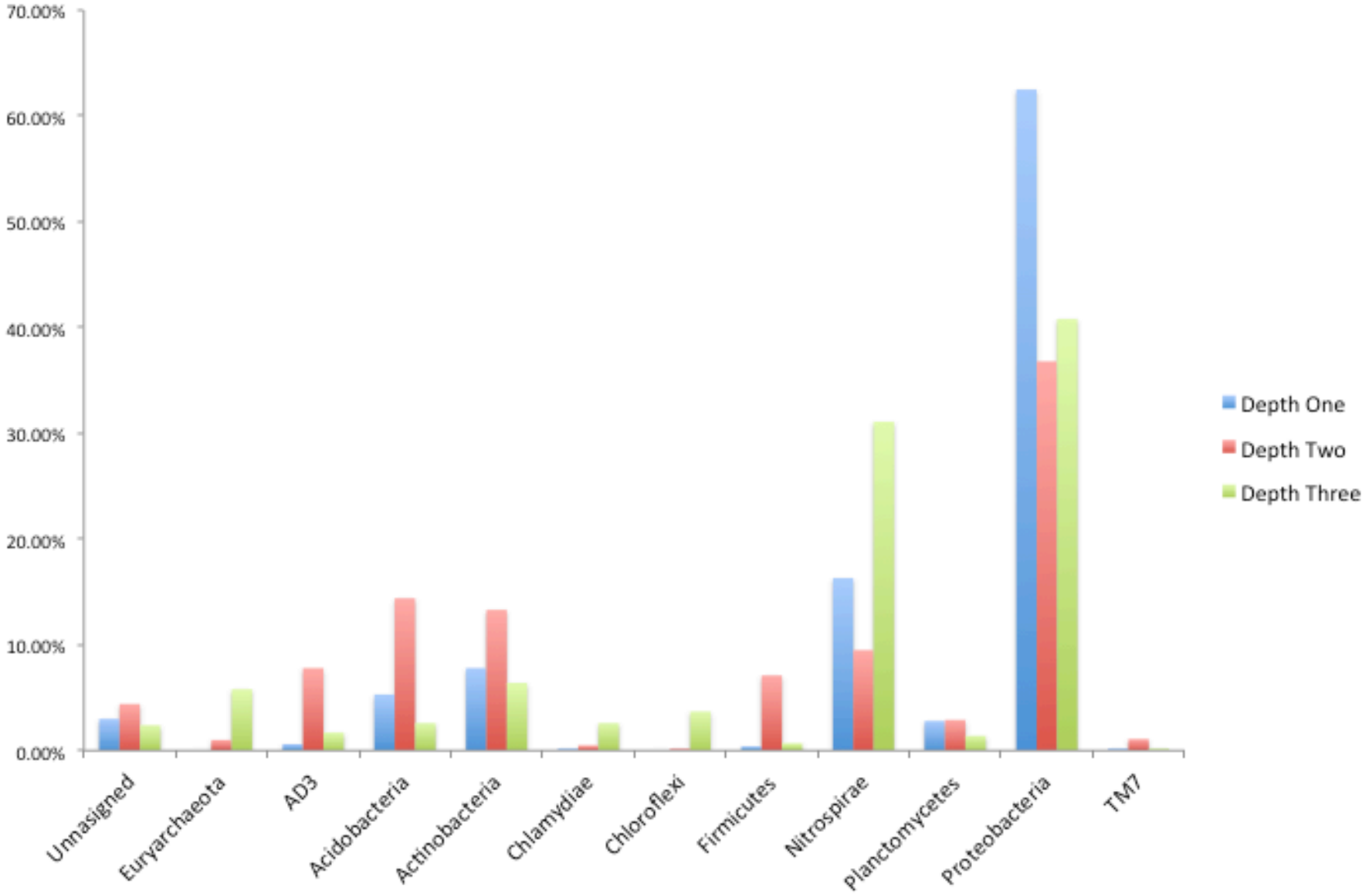


Figure 4.12 Phyla with >1% of reads assigned to them by QIIME from three depths at Wheal Maid site 1.

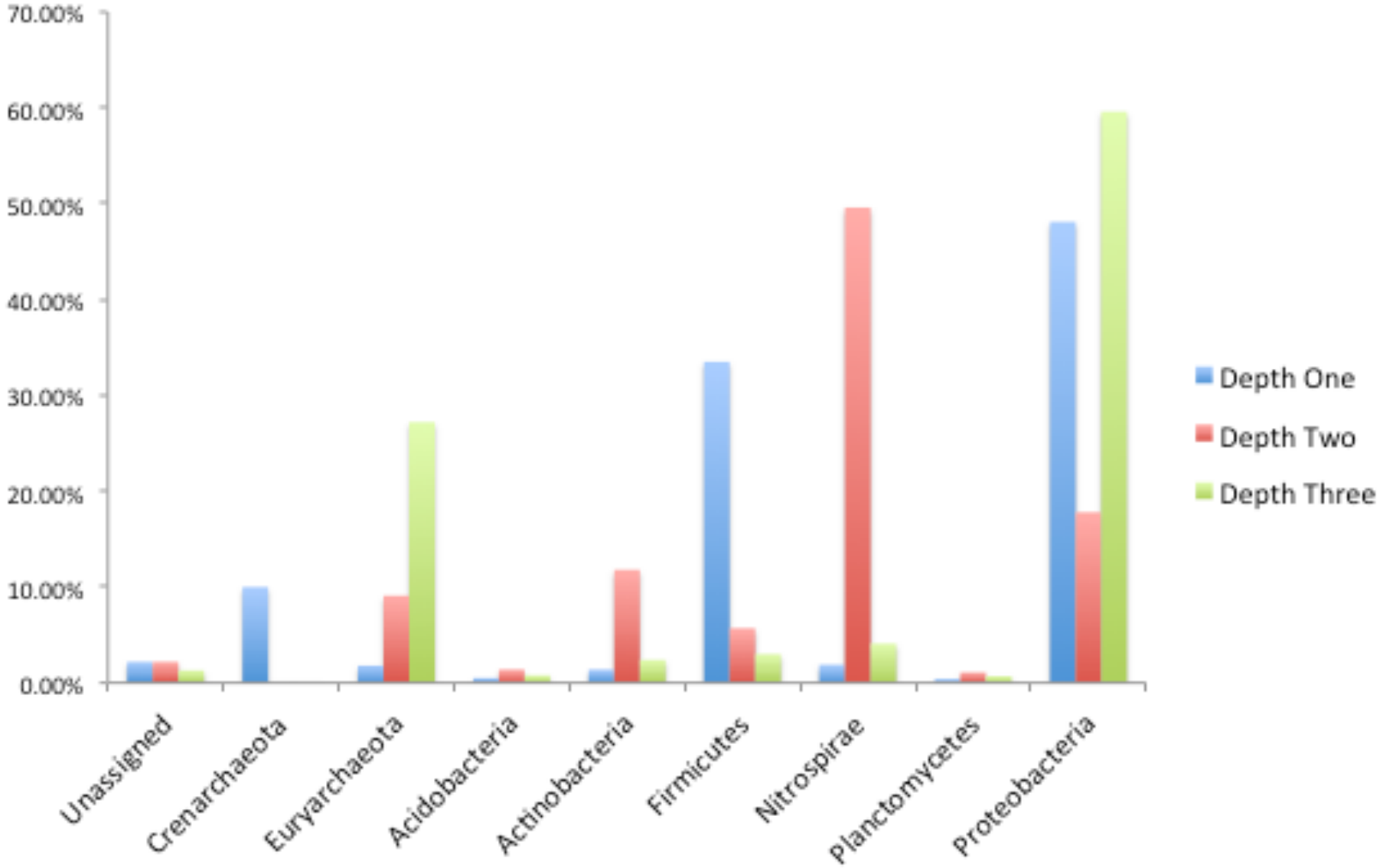


Figure 4.13 Phyla with >1% of reads assigned to them by QIIME from three depths at Wheal Maid site 2.

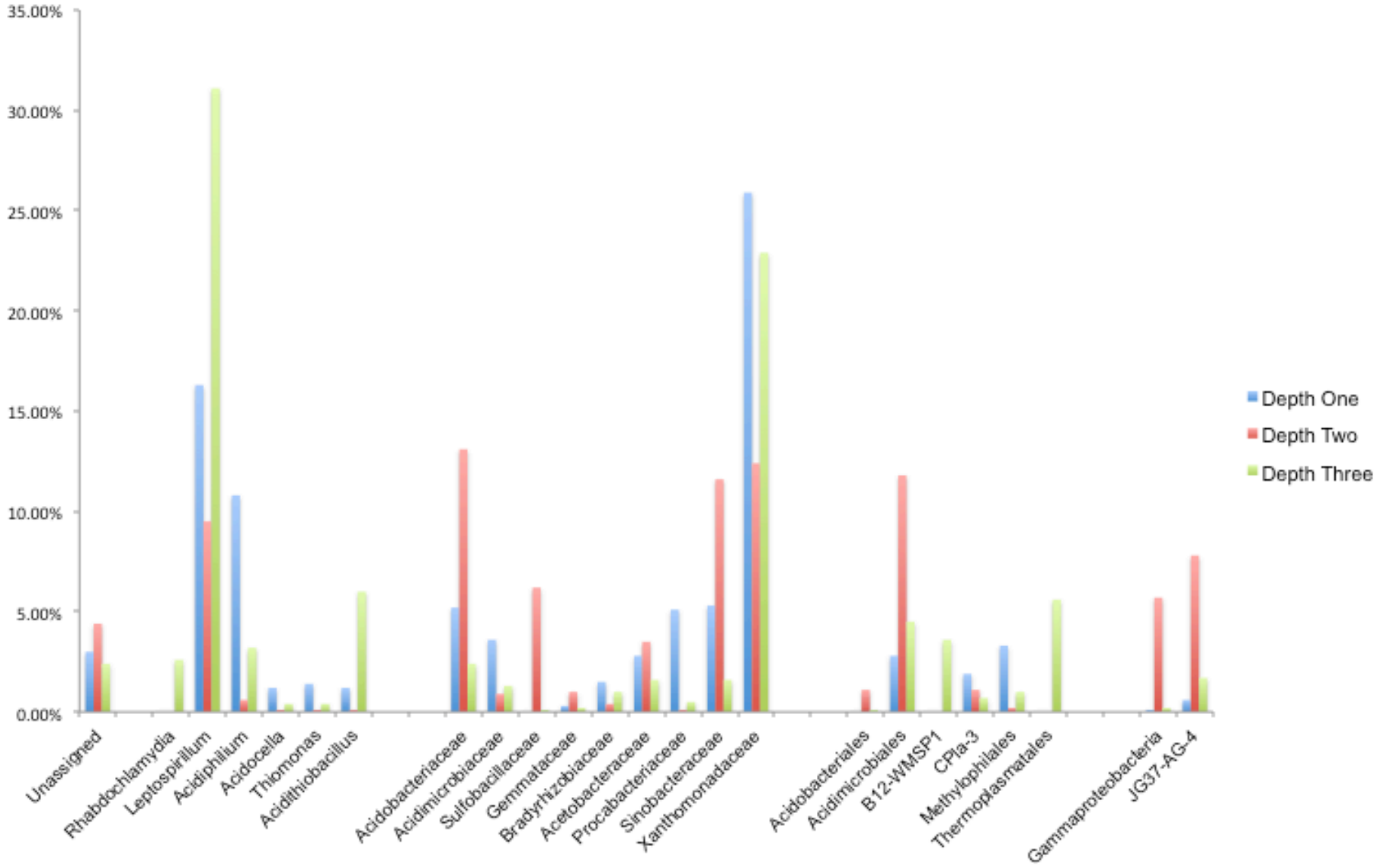


Figure 4.14 Lowest taxonomy reads have been assigned to from three depths at Wheal Maid site 1 using QIIME, taxa with >1% of reads assigned to them at at least one depth are shown.

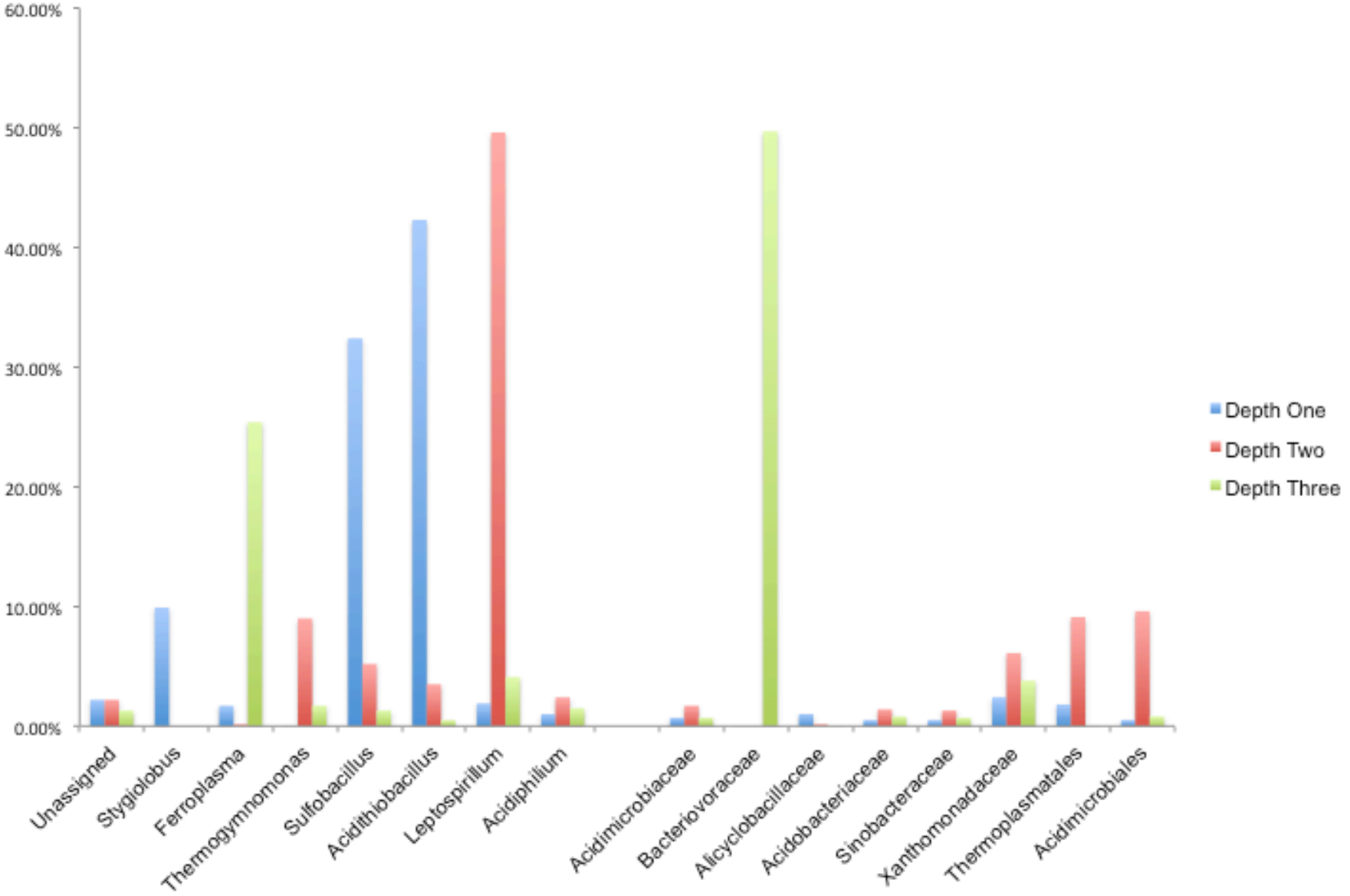


Figure 4.15 Lowest taxonomy reads have been assigned to from three depths at Wheal Maid site two using QIIME, taxa with >1% of reads assigned to them at at least one depth are shown.

At Wheal Maid site 2, depth three 49 % of reads are classified as the family Bacteriovoracaceae (Proteobacteria, Oligoflexia, Bacteriovoracales). Reads that had been assigned to this taxon were used in a BLAST alignment against the NCBI 16S rRNA database and the NCBI nucleotide database. Results from the BLAST alignment as well as members of the Bacteriovoracaceae family were used to create a phylogenetic tree (Figure 4.16) This tree shows these 16S rRNA gene sequences from Wheal Maid, site 2 depth three are most closely related to a number of uncultured bacteria, three of which were obtained from AMD at Iron Mountain, California and one of which was obtained from AMD at TongLing pyrite mine, China. This phylogeny suggests that the large number of reads classified as Bacteriovoraceae are more likely to be acidophilic bacteria that have not been previously classified into a specific taxon.

For Wheal Maid site 1 a large proportion of reads was classified as the family Xanthomonadaceae (Proteobacteria, Gammaproteobacteria, Xanthomondales). As before, a BLAST alignment was performed and phylogenetic tree created using reads assigned to Xanthomonadaceae along with sequences from members of the Xanthomonadaceae family and sequences identified as similar from the BLAST alignment (Figure 4.17). This tree groups nine out of ten of the reads from site 1, depth one together in a clade closest to a number of uncultured bacteria and species of *Metallibacterium* (Proteobacteria, Gammaproteobacteria, Xanthomondales, Rhodanobacteraceae) whilst one read is in the clade with *Metallibacterium* and the uncultured bacteria. The uncultured bacteria present in this phylogenetic tree were obtained from AMD at the Tinto river, a sulfidic mine tailings dump and an AMD site at Carnoules (France). *Metallibacterium* were first isolated from an acidic biofilm in a German pyrite mine **(**Ziegler *et al*., 2013) and since then have been found in a wide range of mine environments where they have been shown to thrive in a range of temperatures and pH levels. It is likely that a number of reads that have been assigned to the family Xanthomonadaceae are actually more closely related to *Metallibacterium.*

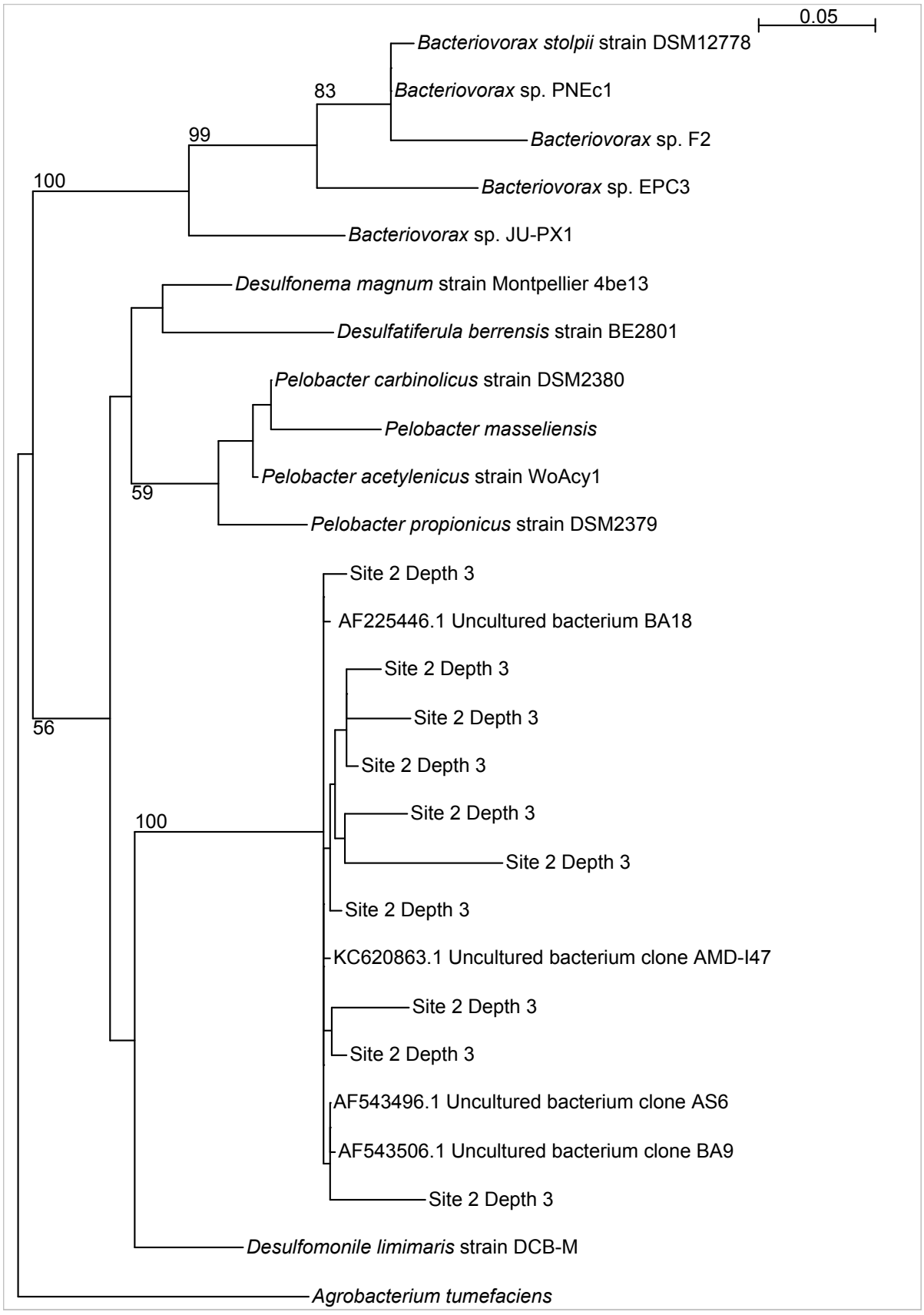


Figure 4.16, Maximum likelihood tree showing reads from Wheal Maid site 2, depth three which had been assigned to the family Bacteriovoracaceae, organisms classed as similar to these reads by BLAST and members of the family Bacteriovoracaceae. The tree is rooted at *Agrobacterium tumefaciens* and bootstrap values over 50 are shown.

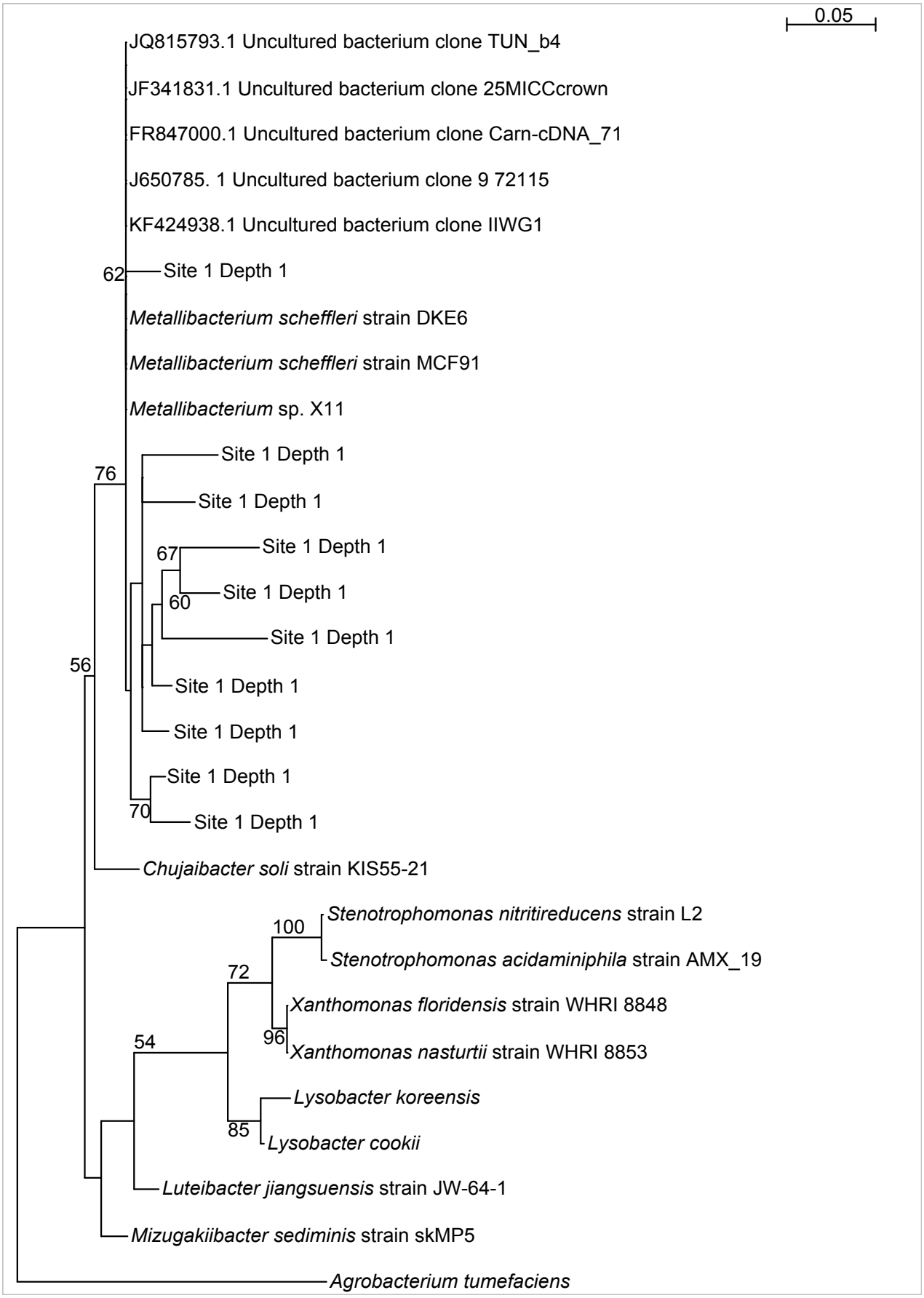


Figure 4.17, Maximum likelihood tree showing reads from Wheal Maid site 1, depth one which had been assigned to the family Xanthomonadaceae, organisms classed as similar to these reads by BLAST and members of the family Xanthomonadaceae. The tree is rooted at *Agrobacterium tumefaciens* and bootstrap values over 50 are shown.

4.4.2 Wheal maid site 1 appears to have a more complex community than site 2.

Wheal Maid site 1 appears to have a more complex community across the three depths than site 2; this can be seen in Figures 4.10 and 4.11 which show graphical representations of the number of different taxa assigned to each sample (to the lowest taxonomic level, down to genus). Figure 4.18 shows alpha rarefaction plots and Figure 4.19 shows Beta-diversity plots. The alpha rarefaction plots demonstrate that the samples with the highest measure of both species richness and diversity are site 1, depth two followed by site 1, depth 1. At both site 1 and site 2 the largest phylum present is Proteobacteria, with 47 % of all reads from site 1 and 42 % of all reads from site 2 assigned to this phylum (Figures 4.12 and 4.13). However, when looking at each depth separately, the largest phylum at site 2 depth two is Nitrospirae, with 49 % of reads assigned to it. Reads assigned to the phylum Nitrospirae have been further classified as *Leptospirillum* which is present in all samples; 16 %, 10 % and 31 % of depths one, two and three respectively at site 1, and 2 %, 50 % and 5 % of depths one, two and three at site 2 are assigned to *Leptospirillum* (Figures 4.14 and 4.15). As previously discussed, *Leptospirillum* oxidises iron and is a key organism in the generation of AMD. The large proportion of *Leptospirillum* at site 2, depth two may be as a consequence of the high levels of iron present at this depth (21.9 %) compared to the other depths (7.3 % at depth one and 3.0 % at depth three).

Proteobacteria is the dominant phylum across Wheal Maid site 1. Genera from this phylum with over 1 % of reads assigned to them, include *Acidiphillium*, *Acidocella*, *Thiomonas* and *Acidithiobacillus* (Figure 4.14*).* Both *Acidiphilium* and *Acidocella* are from the family Acetobacteraceae and have previously been isolated from AMD sites. At Wheal Maid site 1 the obligate aerobe *Acidiphilium* is most abundant at depth one (surface) where it is present at 11 %, whilst *Acidocella* is also most dominant at the surface but only at levels of 2 %. However, there are also 5 %, 13 % and 2 % of reads from the three depths respectively which have only been classified to the Acidobacteriaceae level. The Proteobacteria genus *Thiomonas* is present at site 1 at levels just over 1 % but only at the surface level. Members of this genus have previously been found in AMD where some are able to oxidise arsenic. Site 1, depth one has the

highest levels of arsenic across both sites and it is therefore possible that this strain of *Thiomonas* is utilising this arsenic. *Thiomonas* was not found across site 2 at any depths. Some members of *Thiomonas* were reclassified into the genus *Acidithiobacillus* in 2000 (Kelly *et al*., 2000), which is also present at site 1 at depths one and three at 1 % and 6 % respectively. *Acidithiobacillus* spp. are acidophilic obligate autotrophs and frequently found in AMD. *A. ferrooxidans* and *A. thiooxidans* are known to be important in the generation of AMD as they oxidise ferrous iron and sulphur respectively (Harneit *et al*., 2006). Site 1 also has many reads assigned to the phylum Proteobacteria which have only been further classified to the family level; these families are: Bradyrhizobiaceae, Procabacteriaceae, Sinobacteraceae, Xanthomonadaceae and Acetobacteraceae. Xanthomonadaceae have been discussed above and are likely to be a misclassification of bacteria related to *Metallibacterium*, and Acetobacteraceae have also been discussed above. Procabacteriaceae, members of which have been previously identified as an obligate endosymbiont of *Acanthamoeba,* is present in depth one with 5 % of reads assigned to this family. Endosymbionts of *Acanthamoeba* have previously been isolated from AMD environments and sequences related to *Acanthamoeba* have been found in a study looking at the microbial population of AMD in the Rio Tinto (Amaral-Zettler *et al*., 2011; Baker *et al*., 2003). Finding Procabacteriaceae at the surface of Wheal Maid site 1 indicates *Acanthamoeba* may be present here. Sinobacteraceae are present at 5 % , 12 % and 2 % across the three depths at site 1. Sinobacteraceae which could not be assigned to any described species have previously been isolated from sediment at the Rio Tinto river and AMD barrens in the USA (Sánchez-Andrea *et al*., 2011; Rojas *et al,* 2016).

Like Wheal Maid site 1, site 2 contains a number of Proteobacteria (Figure 4.13), however the range is not so diverse (Figure 4.15). *Acidithiobacillus* is dominant at site 2, depth one where it is present at 42 %, with 4 % of reads also assigned to this genus at depth two. *Acidiphilium* is also present at site 2 but at lower percentages than site 1 with between 1–2 % of reads assigned to this genus across the three depths. Sinobacteraceae is also present, but only at depth two at just over 1 %. The most abundant member of the Proteobacteria at site 2, depth three is Bacteriovoracaceae with 50 % of reads assigned to this

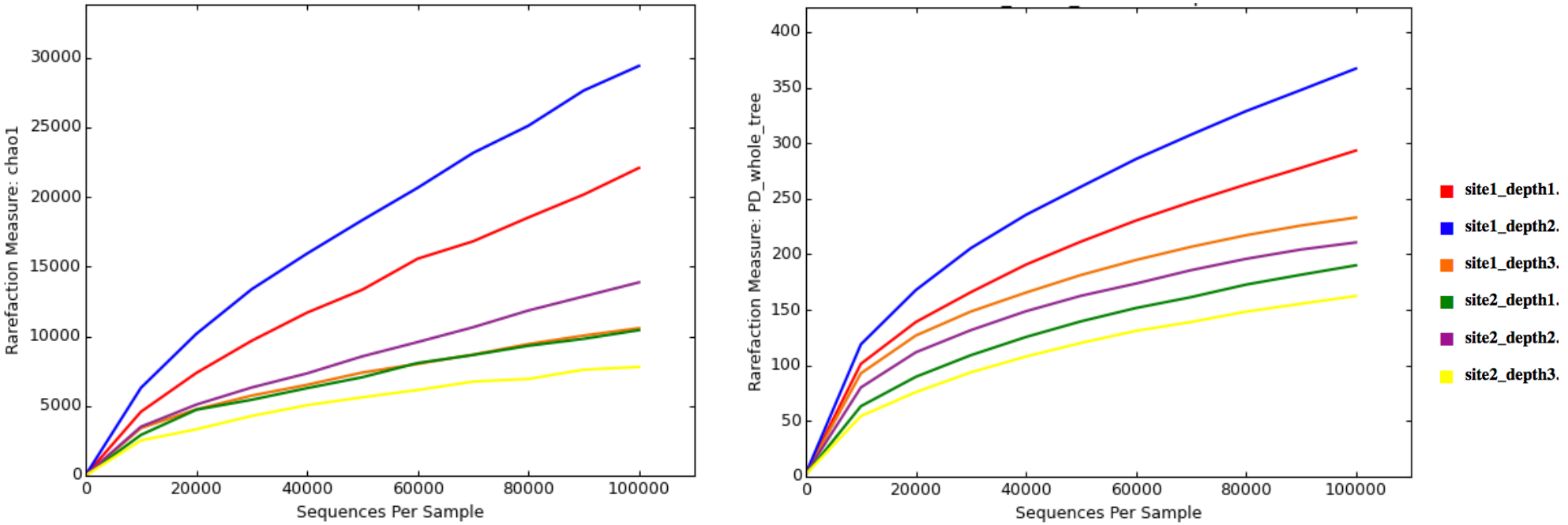
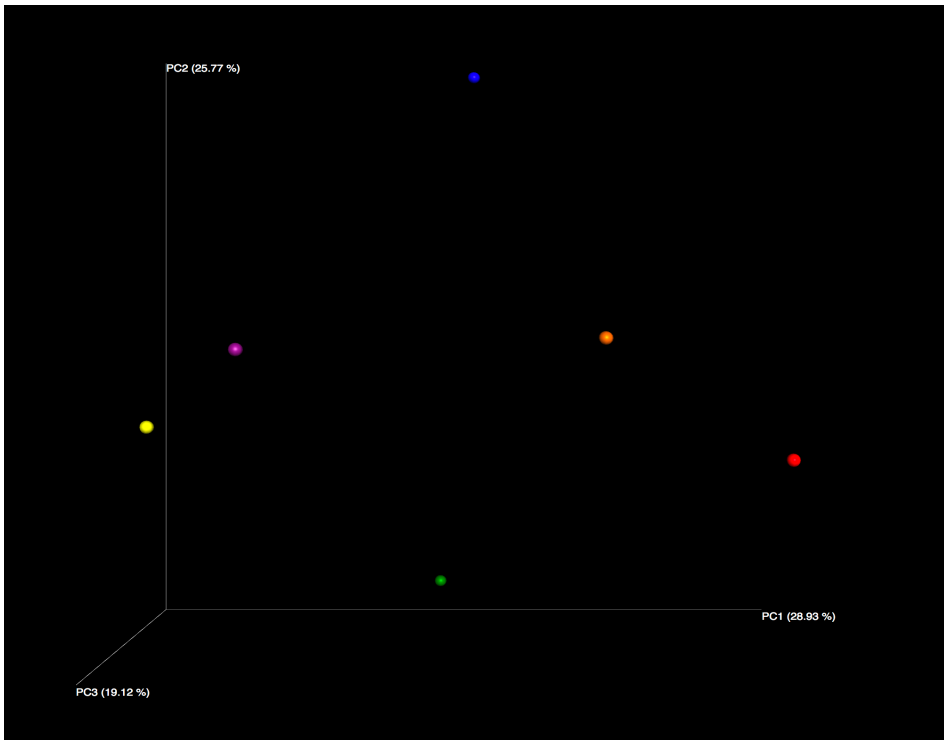
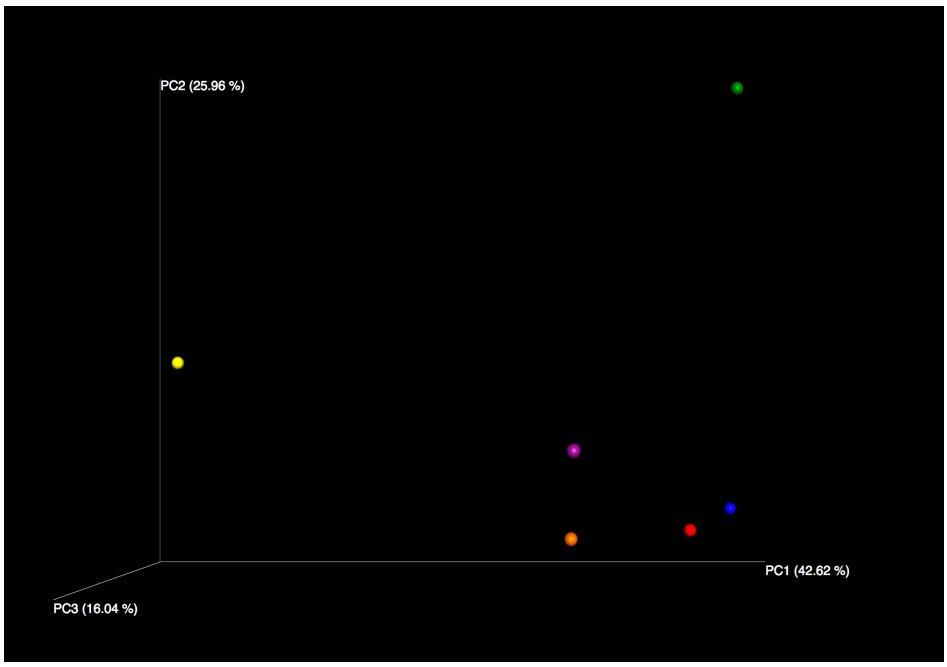
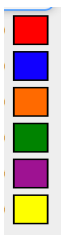


Figure 4.18 Rarefaction curves on the species richness (Chao1) and diversity (PD whole tree) of sediment at Wheal Maid sites 1 and 2, at three sediment depths. A sampling depth of 100000 has been used.







Site 1, depth 1
Site 1, depth 2
Site 1, depth 3
Site 2, depth 1
Site 2, depth 2
Site 2, depth 3

Figure 4.19 phylogentic beta diversity (unweighted UniFrac, top) and non-phylogentic beta diversity (weighted UniFrac, bottom) plots of sediment at Wheal Maid sites 1 and 2, at three sediment depths.

family, however, as previously discussed, this is likely a misclassification of novel acidophiles. As noted above, Acidobacteria is present at Wheal Maid site 1, with reads from this phylum further classified to the family Acidobacteriaceae. (5 % ,13 % and 2 % of reads from depths one, two and three respectively. At site 2, however, only 1 % of reads are assigned to Acidobacteriaceae. Acidobacteriaceae are heterotrophic acidophiles commonly found in AMD-impacted environments.

Ferroplasma is the most abundant archaeon at site 2, with 25 % of reads from depth three and 2 % of reads from depth one assigned to this genus. Depth two has 9 % of reads assigned to the order Thermoplasmatales to which *Ferroplasma* belongs. Site 1 has no reads assigned to the genus *Ferroplasma*, but 5.6 % of reads from site 1, depth three are assigned to Thermoplasmatales. Species of *Ferroplasma* are acidophilic and are often found in acid mine drainage where they oxidise iron and accelerate the rate of disintegration of metal sulphide minerals, contributing to the formation of AMD (Edwards *et al*., 2000).

4.5 Summary

This chapter has used 16S rRNA gene sequences to assess the complexity of the microbial population at AMD locations Wheal Jane and Wheal Maid. The differences in the microbial populations at these two locations have been highlighted, but further sampling would be required to draw definitive conclusions as to the Wheal Jane population due to the changeable nature of the site. Organisms identified as living in Wheal Maid sediment have included those which are frequently found in and which play a key role in the formation of AMD, including *Leptospirillum*, *Acidiphilium* and *Acidithiobacillus,* but there are differences in the populations depending on both the depth and location of the sediment samples taken from this location.

This study has highlighted the potential for misclassification of 16S rRNA gene sequences by the two different classifiers used (Kraken and QIIME) and the importance of ensuring the classifications suggested are correct. This appears to be a problem for novel species which may not be present in the databases used for comparisons, resulting in classification tools classifying novel reads as a similar species present in databases.

Although the use of 16S rRNA gene sequence analysis has proved useful in gaining a greater understanding of the range of microorganisms present in Wheal Maid and Wheal Jane AMD, there are limitation to the information that can be gained through 16S rRNA analysis alone; classification to the species level is not possible and there is no information regarding metabolic pathways. A metagenomic study of the microbial communities present is required if a full understanding of the microbial population is to be achieved; metagenomics allows for taxonomic classification down to the species level and, essentially, allows for full annotation of genes, enabling bioremediation or other pathways of interest to be identified. Chapter 5 will use metagenomics to address this, focussing on the microbial population present in AMD at Wheal Maid.

Chapter five: Metagenomic analysis of the microbial community found in acid mine drainage at Wheal Maid

5.1 Introduction

Chapter 4 addressed questions regarding the complexity of the microbial community found in sediment at the Wheal Maid tailings lagoon, Cornwall. As previously discussed, this site is an acidic environment contaminated with a range of toxic metals. Chapter 4 determined, through the use of 16S rRNA gene analysis, that a range of bacteria and archaea are living in Wheal Maid sediments, most of which are typical of AMD environments. The complexity of the community differed slightly depending on location and depth of the sediment sample. Although chapter 4| allowed for an initial glimpse into the microbial community at Wheal Maid it also had its limitations; organisms could only be identified as far as genus level and many reads were only identified to taxonomic levels higher than genus. Furthermore, the use of 16S rRNA gene analysis offers no insights into the functions that the microbial population may be carrying out and cannot fully answer the question of whether there are microbes with the potential for bioremediation present. To fully understand the taxonomy and function of the microbial community full genomic sequence data is needed.

5.1.1 Metagenomics for the study of microbial communities found in AMD

As previously discussed (Chapter 1), the identification of microbial life found in AMD has previously been investigated at a number of mine sites across the world. Many bacteria and archaea have been identified that can thrive in the low pH and sulphate/metal-rich environment of AMD contaminated waters and sediments. Metagenomics allows for metabolic pathways and genes of interest to be identified within a microbial population. When studying populations living in AMD there are a number of genes of interest that contribute to the delicate balance of the microbial ecosystem as well as those which contribute to the bioremediation of AMD. Limited levels of organic carbon and nitrogen in AMD means organisms capable of carbon and nitrogen fixation coupled with iron and sulphur oxidation are required in AMD environments to maintain a steady supply of these elements within the ecosystem (Chen *et al*., 2016). Genes related to survival in the harsh environmental conditions are a common feature of organisms found in AMD and novel metal and acid resistance genes have

previously been discovered through metagenomic analysis of AMD microbial communities (Guazzaroni *et al*., 2013; Mirete *et al*., 2007). Sulphate-reducing bacteria (SRB) have previously been discussed with regard to their potential for the biological treatment of AMD (Chapter1); discovering acidophilic SRB which have high levels of resistance to the metals associated with AMD and which could therefore be utilised in AMD bioreactors is an area of high interest to those developing bioremediation technologies (Martins *et al*., 2009). Within bioreactors a key process that takes place is the degradation of organic matter which is then utilised by SRB, however, the microbes that carry out this task are largely unstudied; metagenomics may identify such microbes that are able to live alongside bioremediating bacteria, providing crucial nutrients without directly being involved in bioremediation processes.

5.1.2 Aims of this chapter

This chapter aims to carry out a molecular microbiology study on sediment samples taken from two sites and two depths at the Wheal Maid tailings lagoon using shotgun metagenomic sequencing. Analysis of metagenomic sequence data from Wheal Maid will be carried out with the aim of identifying microorganisms living in AMD-contaminated sediments as well as identifying genes of interest related to how microorganisms present survive in the harsh environmental conditions.

5.2 Materials and methods

5.2.1 DNA extractions and sequencing

Sediment samples from Wheal Maid were obtained by Chris Bryan *et al.* of the University of Exeter Environment and Sustainability Institute, Penryn Campus. Samples were taken from two locations at Wheal Maid (Chapter 4, Figure 4.1). Three cores were taken at each location and 8 replicates were taken from the surface and 30 cm depth of each core. Bryan *et al*. carried out DNA extractions from sediment samples using the Mo Bio PowerSoil DNA isolation kit. Genomic libraries were prepared using the Nexetera XT DNA library prep kit and 300 bp paired end sequencing was carried out on an Illumina MiSeq by the Exeter Sequencing Service.

5.2.2 Bioinformatics tools and software

Table 5.1 shows software databases and websites used in this study. Default parameters were used for all programmes unless stated otherwise below.

5.2.3 Assembly and taxonomic classification of sequences

Sequences were trimmed using Trim Galore, and assembled using MetaSPAdes. Quast was used for assessing assemblies.

Kraken was used for taxonomic classifications using both FASTA and FASTQ input and the standard Kraken database. Pavian was used for visualisation of Kraken-assigned taxonomy.

5.2.4 Contig binning

The following Anvi'o workflow was run on a FASTA file of all samples assembled together (using MetaSPAdes):

anvi-gen-contigs-database

anvi-run-hmms

anvi-run-ncbi-cogs

anvi-import-taxonomy

Individual samples were aligned against the FASTA file of all samples assembled together using BWA mem; the resulting SAM files were converted to BAM files using SAMtools. The following Anvi'o workflow was then run on each BAM file:

anvi-profile

anvi-merge

anvi-interactive

anvi-summarize

5.2.5 Functional annotation

Nucleotide sequences were translated into amino acid sequences using TranSeq and uploaded to GhostKoala for annotation with KO numbers.

FMAP was run using the following workflow:

FMAP_download.pl

FMAP_mapping.pl

FMAP_quantification.pl

FMAP_table.pl

FMAP_comparison.pl

Whole genomes binned using Anvi'o were uploaded to RAST for annotation.

A cladogram of RecA sequences was generated using muscle for alignment of sequences and PhyMLfor tree building, from the SEAVIEW package.

Table 5.1 Bioinformatic software and websites used in this study.

Name	**Version**	**Available from:**	**Reference**
Anvi'o	4	http://merenlab.org/2016/06/26/installation-v2/	Eren *et al*., 2015
DeconSeq	0.4.3	http://deconseq.sourceforge.net/	Haque *et al*., 2015
FMAP	1	https://github.com/jiwoongbio/FMAP	Kim *et al*., 2016
GhostKoala		http://www.kegg.jp/ghostkoala/	Kanehisa *et al.,* 2016
Kegg	73.1	http://www.kegg.jp/kegg/	Kanehisa *et al*., 2000
KRAKEN	0.10.6	https://ccb.jhu.edu/software/kraken/	Wood & Salzberg, 2014
One Codex	Accessed January 2018	https://www.onecodex.com/	Minot *et al*., 2015
Pavian	0.6.2	https://github.com/fbreitwieser/pavian	Breitwieser & Salzberg, 2016
Quast	2.3	http://bioinf.spbau.ru/quast	Gurevich *et al.*, 2013
SAMtools	0.1.19	http://samtools.sourceforge.net	Li *et al.,* 2009
SEAVIEW	4.5.3	http://doua.prabi.fr/software/seaview	Gouy *et al*., 2010
Transeq	6.6.0.0	https://www.ebi.ac.uk/seqdb/confluence/display/THD/EMBOSS+Transeq	Chojnacki *et al.,* 2017

5.3 Assigning taxonomy to the Wheal Maid microbial community

5.3.1 Assembling and classifying the metagenomic dataset from Wheal Maid leaves a large proportion unclassified

As with chapter 4, sediment samples were taken from two locations at Wheal Maid. Three cores were taken at each location and 8 replicates were taken from the surface and 30 cm depth of each core. DNA extractions were then carried out and sequencing libraries prepared. However, for all samples taken from site 2 depth there was either insufficient DNA or the library failed. This could be due to very low abundance of organisms living at this depth at this location or there could be environmental conditions which have inhibited library preparation and/or DNA extraction. Further exploration of this site would be required to draw firm conclusions as to why this occurred. Therefore, sequence data was obtained for the surface of sites 1 and 2 and at depth for site 1 only. Replicates were initially assembled individually to assess if there were any major differences between the replicates, which there were not. However, it was decided to assemble replicates (and in some cases cores) together for most analysis due to low coverage present across individual assemblies. Statistics for assemblies using MetaSPAdes are shown in Table 5.2. Figure 5.1 shows output from QUAST for assemblies.

Taxonomic classification was initially carried out using Kraken. However, as only between 7.2 – 13.4 % of reads from the three sites were classified, three other taxonomic classification tools were also used: Megan, One Codex and Kegg. Table 5.3 shows that using Megan and Kegg also resulted in low numbers of reads being assigned to a taxon, whilst One Codex managed slightly higher numbers but only to a maximum of 32.5 %. The fact that all four classifiers failed to assign taxonomy to the majority of reads could be down to a number of reasons including poor quality data, poor quality assemblies or the presence of highly novel organisms. To assess if assembly quality was to blame,

Table 5.2 Statistics for metagenomic sequence data taken from Wheal Maid

	No. of raw reads	No. of contigs after MetaSPAdes	Largest contig	Total length	N50
Site 1 Surface	6 522 294	1 252 364	94 789	378 638 747	1005
Site 1 Depth	2 699 018	366 288	298 254	150 295 329	1113
Site 2 Surface	9 969 715	383 758	312 438	171 189 898	1845
All samples	19 191 027	961 408	791 371	669 244 831	1100

Table 5.3 Percentage of the Wheal Maid metagenomic dataset classified using four methods

Method used for classification	Site 1 surface % classified	Site 1 depth % classified	Site 2 surface % classified	All samples assembled together % classified
Megan (Diamond)	4.1	5.2	4.6	6.3
Kraken	7.2	8.7	13.4	14.5
One Codex	27.3	19.4	32.5	21.8
Kegg	7.2	9.1	8.71	10.6

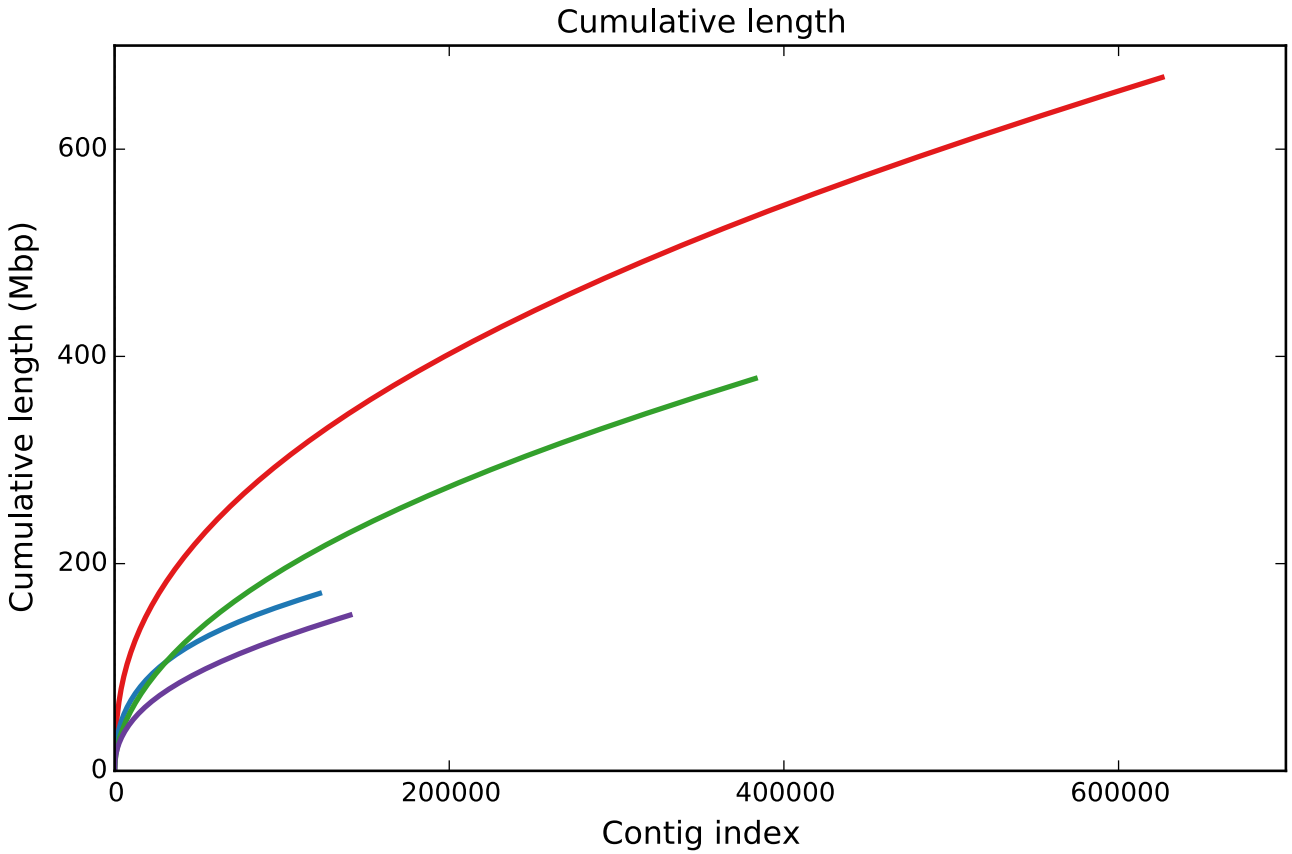


Figure 5.1 Cumulative length plots for all metagenomic sequence assemblies. Red = all samples assembled together, Green = Site 1, surface; Purple = Site 1, depth; Blue = Site 2, surface.

Kraken was run on reads in both raw FASTQ format as well as assembled FASTA files, with no significant change in the numbers being classified (The percentage classified only differed between 1-2 %). Reads from all samples were also assembled as one individual assembly to achieve maximum coverage and run through all four classifiers, however this only resulted in a small increase in the number of reads assigned to taxa (Table 5.3). Raw FASTQ reads were aligned back against the FASTA file composed of all reads assembled together; 95.9 % of reads mapped, however mean coverage was low at 11.6. This low coverage could be contributing to the poor level of taxonomic classification.

Human contamination of reads was checked for using DeconSeq and was not found to be present. Viral sequences within the assembly of all samples were identified using VirFinder, which classified 237 sequences as phage or prophage. Of these 237 sequences, 196 of them had been unclassified by Kraken.

The largest unclassified contig was 312 582 bp long with a coverage of 33.8. This contig was used as the query in a search against the One Codex database of 83 000 whole bacterial, viral, fungal, archaeal, and protozoan genomes, which classified the sequence as being from Candidatus *Parvarchaeum acidiphilum.* This archaeon is a member of the ARMAN (archaeal Richmond Mine acidophilic nanoorganisms) group, first found in 2006 living in AMD at the Richmond Mine, Iron Mountain, California (Baker *et al.*, 2006). Two phyla comprise the ARMAN group: Micrarchaeota and Parvarchaeota (Castelle *et al*., 2015). ARMAN are filterable archaea with cytoplasmic volumes approaching the proposed minimum required for a free-living, independent lifestyle and have a genome size of only around 1 Mb (Baker *et al.*, 2010). Previous studies of the Richmond mine using 16S rRNA gene sequencing had overlooked the ARMAN group as they contain several mismatches with commonly used 16S rRNA PCR primers, and all archaea discovered from this site had previously been classified as belonging to the order Thermoplasmatales (Baker *et al.*, 2010). Whole genome sequencing of three ARMAN lineages found them to have unusual features, with no biological function inferred for up to 45 % of genes and no more than 63 % of the predicted proteins assigned to a set of archaeal clusters

of orthologous groups. ARMAN are likely to be metabolically dependent on other members of the microbial community, with a 2017 study suggesting they form symbiotic relationships with the acidophilic archaea *Cuniculiplasma divulgatum, however* an understanding of their interactions with other organisms is still not fully understood (Golyshina *et al*., 2017). Since their initial discovery, organisms closely related to ARMAN have also been found, through metagenomic studies, in other locations including the AMD polluted river Rio Tino (SW Spain) and an acidic biofilm at the Ehrt pyrite mine, Germany (Méndez-García *et al*., 2015).

Following on from this analysis of a single unclassified contig, all unclassified reads (in unassembled FASTQ format) were queried against the One Codex database. 26.3 % of the reads had *k*-mer matches with acidophilic archaea from the genera *Ferroplasma*, *Ferroplasmaceae*, *Thermoplasmatales*, *Cuniculiplasma*, as well as the ARMAN Candidatus *Parvarchaeota archaeon* and Candidatus *Parvarchaeum acidophilus*. This would suggest that small quantities of archaea are present, some of which are likely to be ARMAN-like organisms. These archaea may not be classified with previously used tools due to being present in low numbers and in very fragmented form. Less than 1 % of unassigned reads aligned against the whole genome sequence of Candidatus *Parvarchaeum acidiphilum* (with an ANI of 81.65 %) suggesting that a number of archaea in the Wheal Maid community are ARMAN-like, but not previously documented strains.

5.3.2 Taxonomy assigned to Wheal Maid shows differences at each site, but typical AMD species are dominant at both.

As discussed above, a large proportion of reads could not be taxonomically classified by several different methods. However, it is possible to look at the proportion that were successfully classified; figures 5.2-5.4 show taxonomic classification by Kraken for site 1 surface, site 1 depth and site 2 surface.

Proteobacteria is the most abundant phylum at all three sites, with 53.3 % of reads from site 1 surface, 47.5 % of reads from site 1 depth and 67.8 % of reads from site 2 all assigned to it. At site 1, surface and depth, reads have been further assigned to a number of species, all at relatively low abundance,

with the most abundant organism at the species level (*Leptospirillum ferrooxidans*), present at 5.3 and 2.8 % at surface and depth respectively, whilst at site 2 the most abundant organism, (*Acidithiobacillus ferrivorans)*, has had 24.2 % of reads assigned to it. This initial analysis indicates site 2 has dominant organisms living alongside less abundant ones, whilst site 1 has a more even distribution of organisms. This is also in keeping with the 16S rRNA gene sequence analysis carried out in chapter 4, which suggested site 2 was host to a less complex community than site 1.

At Wheal Maid site 1, surface and depth, the most abundant genus is *Leptospirillum.* This is unsurprising; *Leptospirillum* is a typical AMD organism which plays a key role in the oxidation of ferrous iron and is often utilised in the biomining industry (Vera *et al*., 2013). Reads assigned to *Leptospirillum* at site 1 have been further assigned to *L. ferrooxidans* (5.3 % at surface level, 2.8 % at depth) and in smaller numbers to *L. ferriphilum* (0.5 % at surface level, 1.7 % at depth). Both of these species are frequently found in very high numbers within AMD environments and are usually the dominant organism when the temperature within the AMD environment is over 20 °C. *Leptospirillum* was also found in high abundance at site 1 from analysis of the 16S rRNA gene sequence data in chapter 4.

Many other species characteristic of AMD sites are also present at site 1, at both surface and depth, including *Granulicella* spp., *Acidobacterium capsulatum*, *Acidimicrobium ferrooxidans*, *Acidithiobacillus ferrooxidans, Singulisphaera acidiphila*, *Acidiphilium* spp. and *Thiomonas arsenitoxydans.* As well as those typical to AMD environments, a number of organisms not previously documented as living in AMD are present in Wheal Maid sites 1 and 2 including: *Achromobacter* sp., *Terriglobus* spp., *Conexibacter woesei, Rhodopseudomonas palustris, Frankia* spp. and *Rhodanobacter denitrificans.* Interestingly, four of these species are involved in nitrogen-related processes; *Conexibacter* reduces nitrate to nitrite, *Rhodanobacter denitrificans* can perform complete denitrification, *Frankia* is a nitrogen-fixing plant symbiont, and *Rhodopseudomonas palustris* is able to switch between different modes of metabolism depending on environmental conditions, one of which is nitrogen fixation (Monciardini *et al.*, 2003; Prakash *et al.*, 2012; Harriott *et al.*, 1995).

Additionally, as well as utilising nitrogen, the metabolically versatile *R. palustris* can grow with or without oxygen, can use light, inorganic or organic compounds for energy and can fix carbon (Larimer et al., 2004). This would make *R. palustris* a potentially useful member of the AMD community. Although not documented as living in AMD, *R. denitrificans* has previously been isolated from a site contaminated with uranium and other heavy metals and has been shown to thrive in conditions of high nitrate and uranium, and low pH; *R. denitrificans* is therefore of great interest for bioremediation purposes (Prakash *et al.*, 2012).

The most abundant genus at site 2, surface level, is *Acidithiobacillus* with 37.7 % of reads from site 2 assigned to this genus. Site 1 also has reads assigned to *Acidithiobacillus* but only at levels of 4.1 % at surface level and 1.1 % at depth. Two species of *Acidithiobacillus* are assigned to reads from Wheal Maid, site 2: *A. ferrivorans* and *A. ferrooxidans.* The most abundant of the two species at site 2, *A. ferrivorans*, is a facultatively anaerobic, iron and sulphur-oxidising acidophile frequently found in AMD (Hallberg *et al*., 2010). *A. ferrooxidans* also oxidises iron and is a major component of microbial consortia used in bio-leaching; additionally *A. ferrooxidans* can fix CO_2 and nitrogen, making it a primary producer of carbon and nitrogen in acidic, nutrient-poor AMD environments (Valdes *et al*., 2008).

Higher numbers of archaea have been assigned to reads from site 2 (9.5 %), than from reads at either surface level (1.4 %) or at depth (2.5 %) from site 1. This is in keeping with the results in chapter 4, where site 2 had higher levels of archaea assigned to it than site 1. The majority of reads (7.5 %) from site 2 assigned to archaea have been further assigned to *Ferroplasma acidarmanus. F. acidarmanus,* an acidophilic archaeon previously isolated from streams draining iron mines, has high levels of resistance to AMD metals including arsenic, zinc and copper (Dopson & Holmes, 2014). *F. acidarmanus* is a chemolithotroph, deriving energy through the oxidation of sulphur found in iron pyrite and has been found to be the dominant organism in streams with a pH between 0-2 where it often forms biofilms.

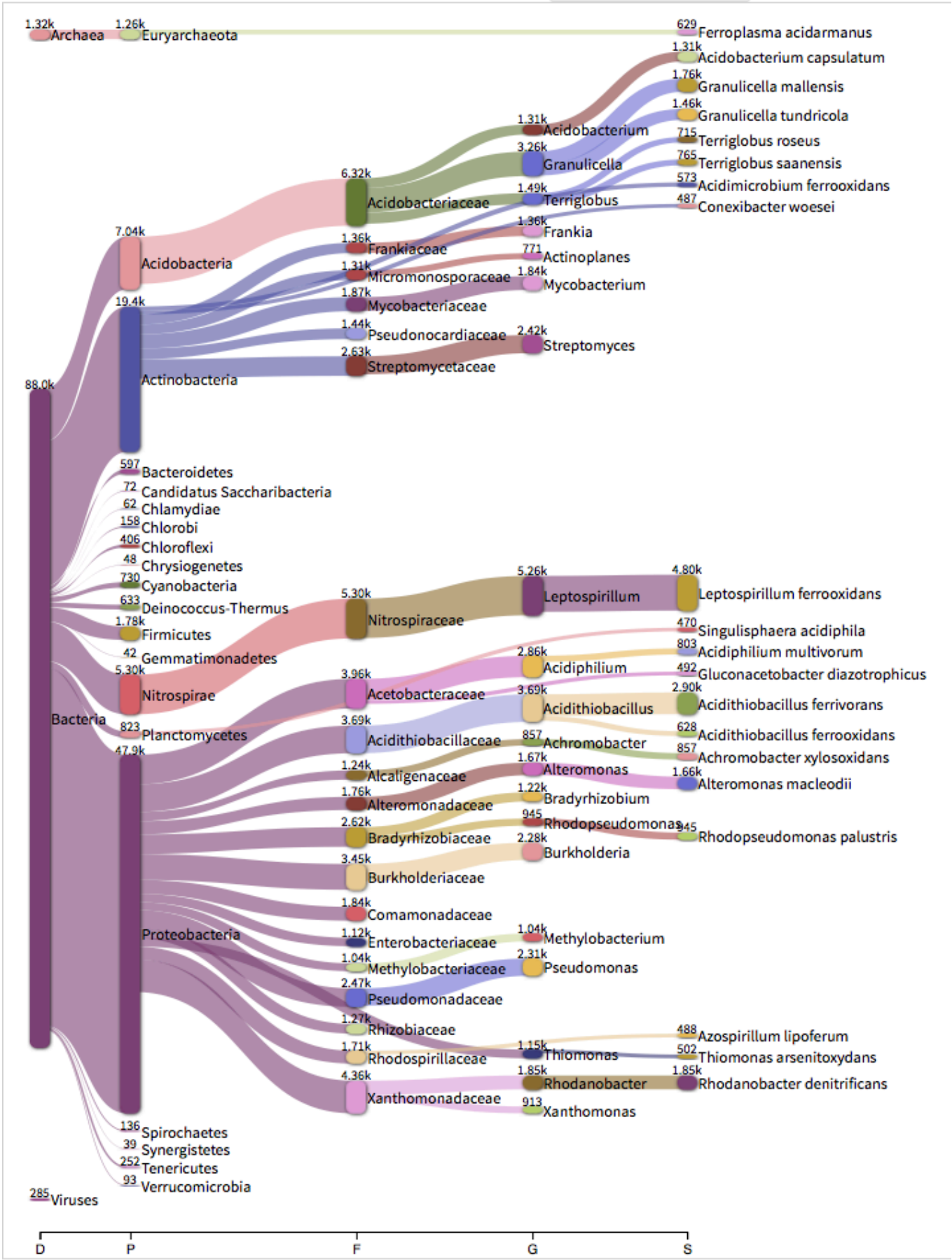


Figure 5.2 Taxonomic classification by Kraken of metagenomic sequence data from Wheal Maid site 1, surface level.

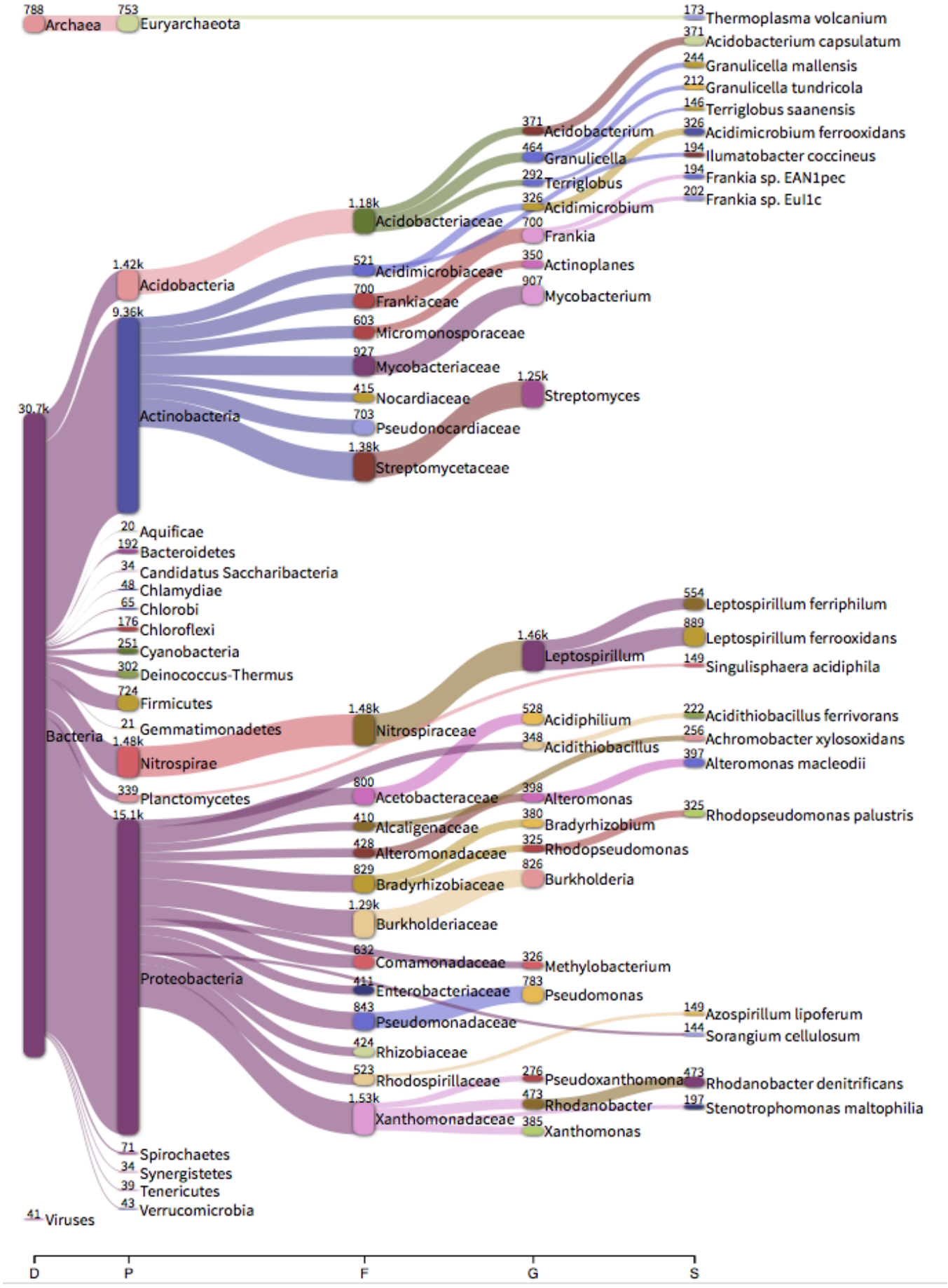


Figure 5.3 Taxonomic classification by Kraken of metagenomic sequence data from Wheal Maid site 1, depth.

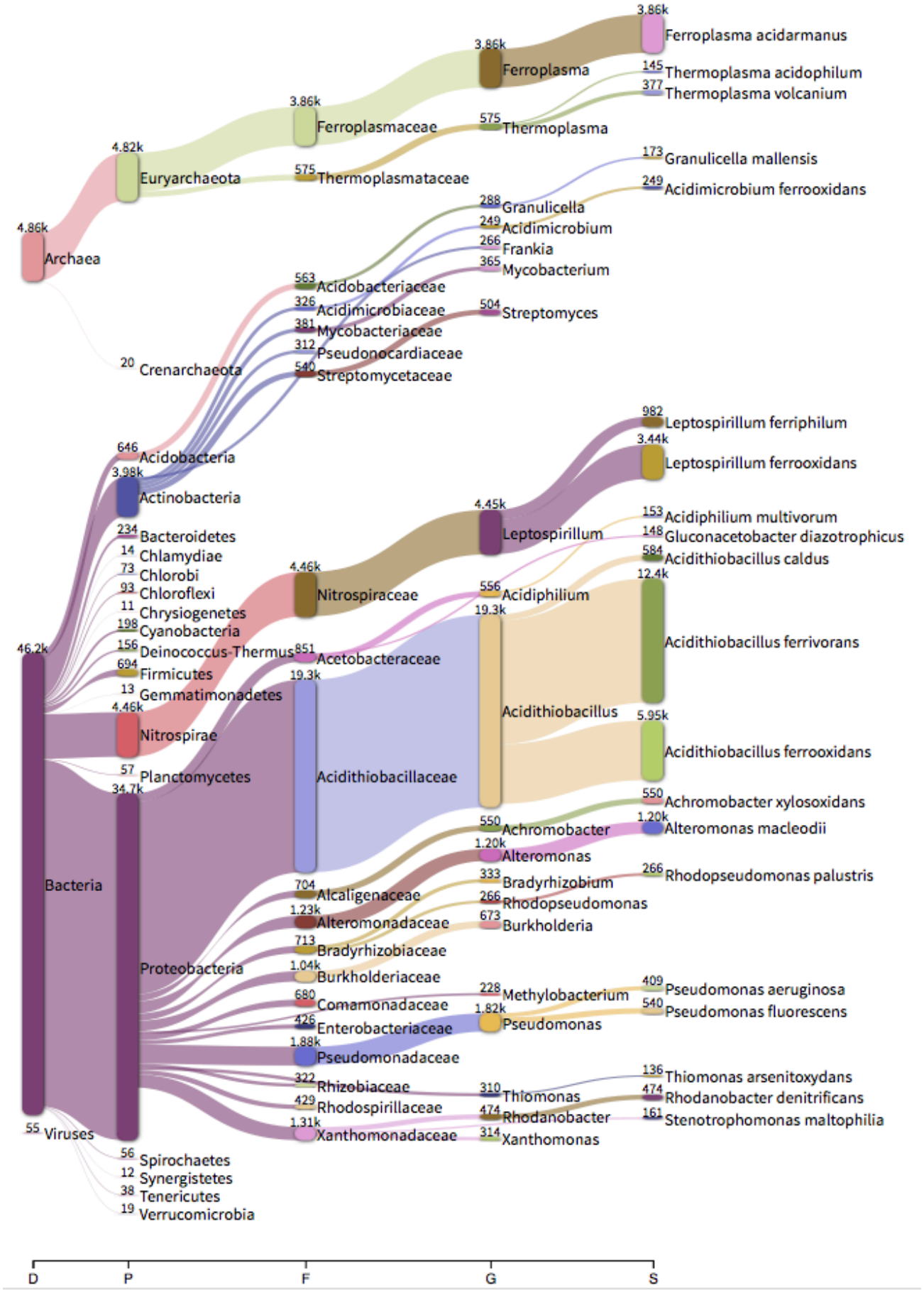


Figure 5.4 Taxonomic classification by Kraken of metagenomic sequence data from Wheal Maid site 2, surface level.

5.4 Constructing whole genomes from Wheal Maid metagenomic sequence data

5.4.1 Binning genomes from metagenomics data

The retrieval of whole genomes from metagenomic data can allow for a greater understanding of microbial ecology, enabling genomic analysis of uncultivable organisms, revealing niche-specific adaptations and allowing for the characterisation of individual taxa present within a larger community (Iverson *et al*., 2012; Sangwan *et al*., 2016). Assembly of short-read metagenomic data (such as that produced by Illumina sequencing) using tools such as MetaSPAdes results in a number of contigs. The reconstruction of individual genomes from these contigs allows for a greater understanding of the function of individual organisms present in the community. Binning genome fragments is one such way this can be achieved. (Alneberg *et al.*, 2014). The binning of individual genomes from metagenomic datasets is a complex task, both in terms of the high computational memory required and the complex biological datasets being worked on. Although advances in this area mean species level genome reconstruction can be achieved, results of genome binning are often dependent on the sequence coverage and on the species diversity present in a metagenomic sample, often resulting in only partial genome reconstruction (Parks *et al*., 2015; Parks *et al*., 2017).

In this study Anvi'o platform software (Eren *et al*., 2015) was used to bin genome fragments, assembled using MetaSPAdes, for sequence data from the two sites at Wheal Maid. Figure 5.5 shows the results of automated genome binning by Anvi'o, which resulted in the generation of 63 genome bins with various levels of completion and redundancy. As well as carrying out automated genome binning, Anvi'o also allows for manual refining of bins by the user, which is especially useful for bins with high levels of redundancy. Figure 5.6 shows genome binning after manual refinement of such bins, which resulted in 81 bins with redundancy levels (dark red bars) lower for the majority of bins

than before refining. Genomes with over 50 % completion are highlighted in blue.

All 30 genomes with over 85 % completion were uploaded to RAST for annotation, and further analysis to identify their taxonomy was carried out (discussed below). Statistics for genomes uploaded to RAST are shown in Table 5.4 from which it will be noted that 22 of these genomes are >90 % complete. Functional annotation is discussed further in sections 5.5-5.7.

5.4.2 Analysis of genomes from Wheal Maid

The identification of housekeeping genes from genomes constructed from the Wheal Maid metagenome was carried out using Anvi'o. Genes were looked for which would be suitable for phylogenetic analysis and taxonomic identification. 16S rRNA gene sequences were only available for 20 genomes, however, *recA* was available for 59 genomes. These *recA* genes were used to generate a cladogram. A heat map was also generated, showing the relative abundance of the genomes from which *recA* was extracted across the different samples. Figure 5.7 shows the cladogram alongside the heat map.

The RAST annotation pipeline establishes phylogenetic context of uploaded genomes by taking a set of protein-coding genes from the genome and using them in a BLAST alignment against sets of corresponding genes in their database. 'Nearest neighbours' are determined this way and used to assist in further annotation of the genome. Anvi'o attempts to assign taxonomy to genomes by importing taxonomic classifications of contigs, generated using Kraken. However, the majority of reads binned into genomes by Anvi'o did not have any taxonomy assigned to them by Kraken. The *recA* gene from each sample was also used in a BLAST alignment against the NCBI nucleotide database, to determine which organism has the highest level of sequence identity. Table 5.5 shows taxonomy/phylogenetic context assigned by these methods to each genome with over 85 % completion, and Table 5.6 shows taxonomy assigned to genomes with less than 85 % completion through *recA* analysis and Anvi'o.

A BLAST search against the NCBI nucleotide database showed high levels of similarity between *recA* extracted from six genomes (Bins 15.1, 15.2, 15.6, 19.2, 56 and 57) taken from the Wheal Maid metagenome and strains of *Cuniculiplasma divulgatum.* The cell-wall-less, acidophilic, mesophilic, organotrophic and facultatively anaerobic archaeon *C. divulgatum* was first discovered in 2016; strains were isolated from acidic streamers at copper mine sites in South-West Spain and North Wales, UK (Golyshina *et al*., 2016). As discussed in 5.3.1, ARMAN are a group of archaea first found in AMD at Iron Mountain, USA, and later detected in other AMD sites. ARMAN are metabolically dependent on other members of the microbial community and a 2017 study determined that an ARMAN-2-related organism, Mia14, could be co-cultured with *C. divulgatum* PM4; evidence of laterally transferred genes from *C. divulgatum* suggested that Mia14 relied on *C. divulgatum* as a host with which genetic material could be exchanged (Golyshina *et al*., 2017). As previously discussed, a number of unclassified reads from the Wheal Maid metagenome are likely to belong to the ARMAN group; additionally *recA* extracted from three genomes (Bins 44, 49 and 34.1) and used in a BLAST search against the NCBI nucleotide database had highest levels of similarity to ARMAN member Candidatus *Micrarchaeota archaeon* Mia14. ANI calculations between these genomes and Candidatus *Micrarchaeota archaeon* ranged from 83.0 to 89.8 %. Four of the six genomes with similarity to *C. divulgatum* (Bins 15.1, 15.2, 15.6 and 19.2) and two of the three genomes with similarity to Candidatus *Micrarchaeota archaeon* Mia14 were found in highest abundance levels at Wheal Maid site 2, surface (Figures 5.6 & 5.7), suggesting conditions there may be best suited to these archaea, making it a good site to use in any further studies carried out to investigate these organisms.

Three genomes (with over 85 % completion) had *recA* sequences with a highest similarity to *Leptospirillum ferrooxidans* C2-3. ANI calculations were computed for these three genomes (Bins 22.1, 22.2, 39) against *Leptospirillum ferrooxidans* C2-3. Bin 22.1 had an ANI of 99.0 % making it the same species as *Leptospirillum ferrooxidans* C2-3, whilst Bins 22.2 and 39 had ANI scores of 90.3 and 83.5 % respectively, meaning they are likely to be from the genus *Leptospirillum* but their species is unknown.

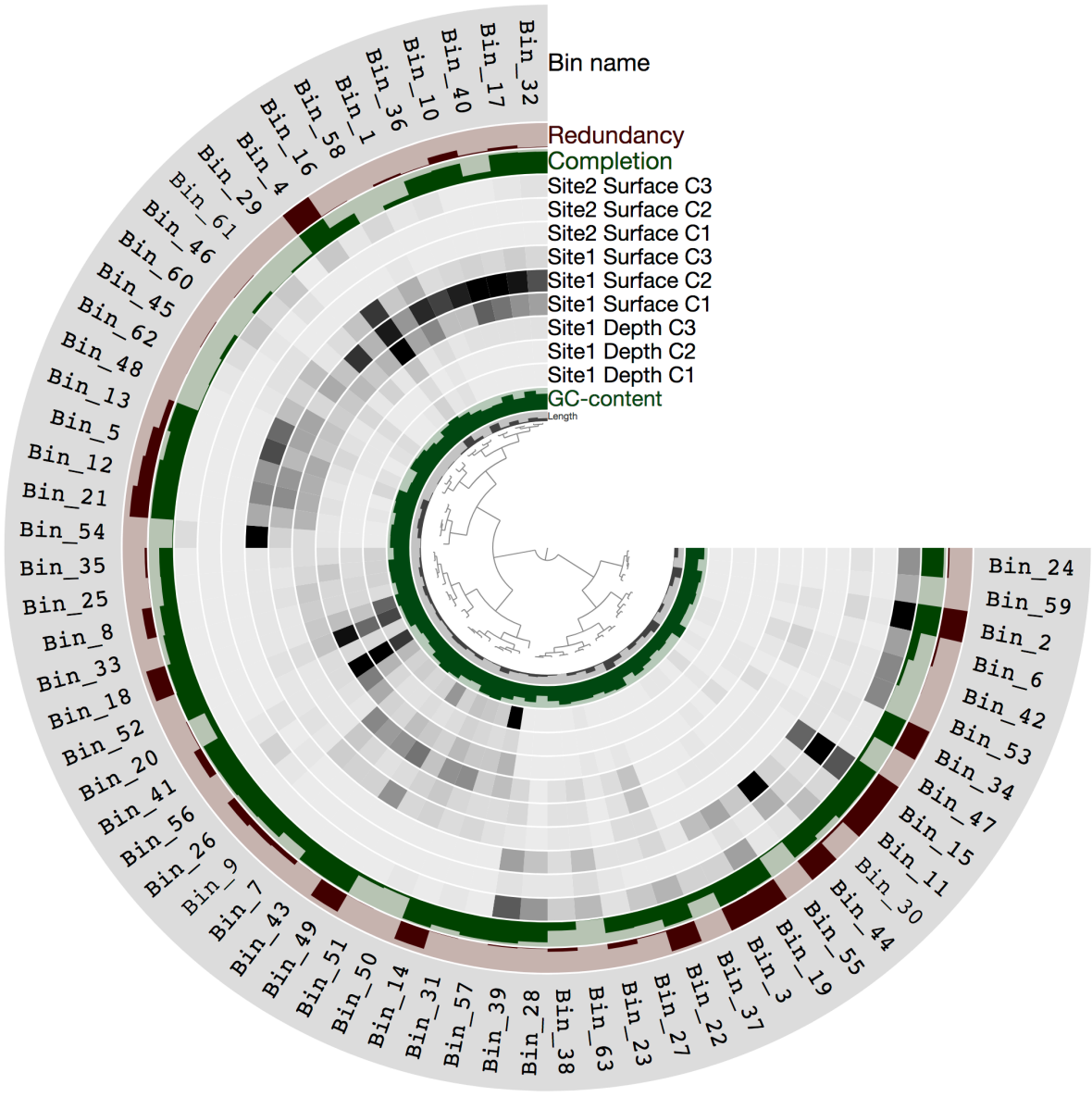


Figure 5.5 Automatic binning of genomes by Anvi'o, showing redundancy, completion and relative abundance at each site. C1=Core 1, C2=Core 2, C3=Core 3.

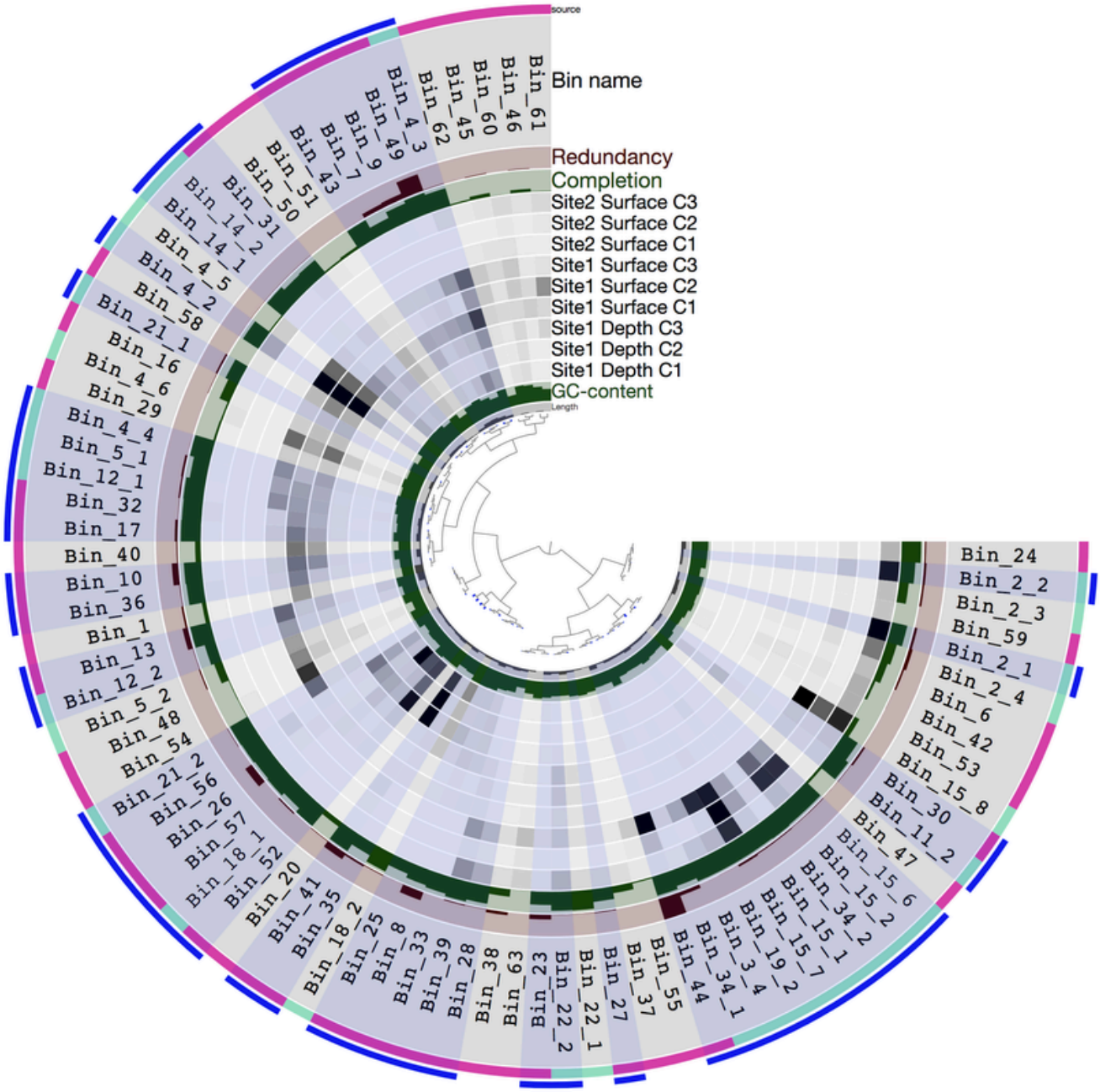


Figure 5.6 Automatic and manual binning of genomes by Anvi'o, showing redundancy, completion and relative abundance at each site. C1=Core 1, C2=Core 2, C3=Core 3. Outer pink and green sections indicate if the genome was binned manually (green) or automatically (pink). The genomes highlighted in blue are over 50 % complete.

Table 5.4 Statistics for genomes (bins) constructed from Wheal Maid metagenomic sequence data with over 85 % completion

BIN	Total length (bp)	Number of contigs	N50	GC content (%)	Completion (%)	Redundancy (%)
3_4	2 282 954	124	33045	62.0	92.8	0.0
4_2	1 656 038	197	11676	67.3	85.6	0.7
4_3	2 081 055	286	9130	66.9	85.6	2.1
4_4	2 926 819	238	17684	67.5	96.4	0.7
11_2	1 685 945	186	13282	42.0	91.9	2.4
12_1	3 620 157	486	9284	53.7	92.0	7.1
13	6 419 408	752	18044	65.3	94.9	24.4
14 _1	2 694 205	321	11515	67.7	94.2	1.4
15 _1	3 041 820	195	32033	38.2	91.9	5.5
15 _2	2 076 028	128	26342	39.9	91.3	2.4
15_6	1 596 100	33	86112	37.2	91.3	3.0
15 7	1 986 776	175	17870	37.3	92.5	10.4
17	4 007 645	697	6733	60.3	92.0	12.9
18_2	1 928 411	185	14175	66.5	88.8	3.7
19_2	2 054 439	59	61260	39.4	93.2	0.6
21_2	2 104 333	148	25362	58.0	94.9	2.8
22 _1	2 984 826	318	15538	50.0	89.9	1.4
22 _2	1 843 808	186	15374	50.3	88.4	1.4
24	3 913 679	462	11712	57.9	89.9	5.7
26	2 034 883	435	4678	44.5	87.6	35.8
32	2 908 366	326	12572	49.6	89.9	2.1
34_2	1 957 519	114	35167	45.3	94.4	4.9
39	2 728 961	236	20584	51.7	92.0	4.3
41	2 869 711	201	24213	44.8	94.4	32.1
43	1 870 510	199	15581	61.9	96.4	0.7
44	4 042 214	26	384693	44.1	90.7	232.1
49	3 264 169	94	56725	47.8	96.9	80.8
52	2 825 127	48	89383	60.1	99.2	0.0
56	1 643 241	18	379273	38.6	92.5	0.0
57	1 486 959	65	43307	42.3	91.9	0.0

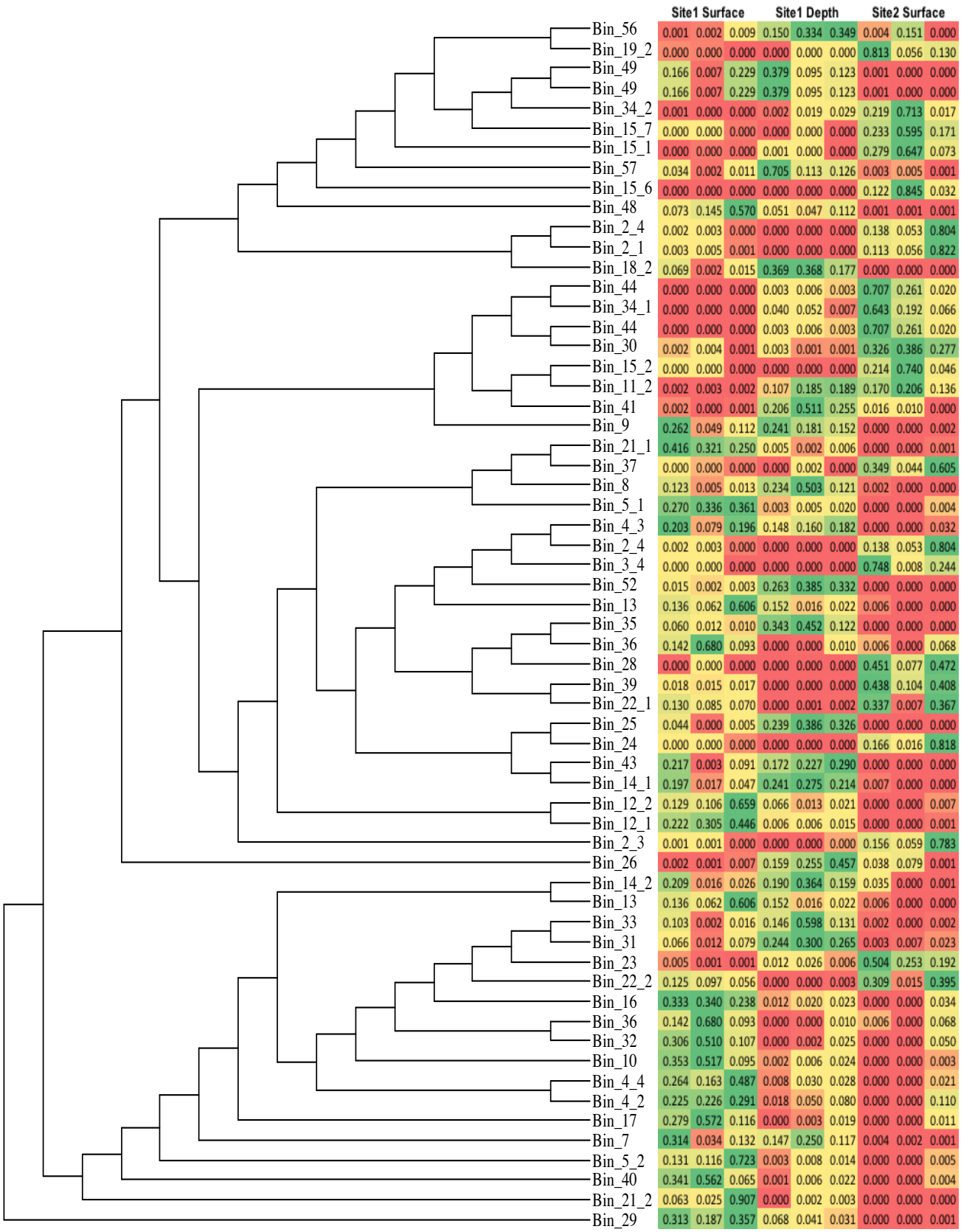


Figure 5.7 Cladogram constructed from *recA* sequences extracted from genomes constructed from Wheal Maid metagenomic sequence data, alongside a heatmap showing abundance of the genomes at Wheal Maid sites 1 (surface and depth) and 2 (surface). Three replicates are shown at each site/depth.

Table 5.5. Taxonomic information for genomes constructed from Wheal Maid metagenomic sequence data (with > 85 % completion).

Genome	Nearest neighbour (RAST)	Genus (Anvi'o)	*recA* BLAST analysis			
			Species	% query coverage	E value	% identity
3.4	*Thioalkalivibrio* sp. HL-EbGR7	None	*Thiohalobacter thiocyanaticus*	90	0	79
4.2	*Dyella japonica* A8	None	*Dokdonella koreensis* DS-123	94	0	88
4.3	*Rhodanobacter spathiphylli* B39	None	*Dokdonella koreensis* DS-123	94	0	88
4.4	*Xanthomonas albilineans*	*Rhodanobacter*	*Dokdonella koreensis* DS-123	95	0	87
11.2	*Ferroplasma acidarmanus*	None	Uncultured archaeon clone ASS_A1	97	0	77
12.1	*Planctomyces limnophilus* DSM 3776	None	*Phycisphaerae bacterium* ST-NAGAB-D1	75	2e-82	69
13	*Thioalkalivibrio* sp. HL-EbGR7	None	*Steroidobacter denitrificans* DSM 18526	93	0	82
14.1	*Ktedonobacter racemifer* DSM 44963	None	*Thermopolyspora flexuosa* NBRC 14349			
15.1	*Ferroplasma acidarmanus*	None	*Cuniculiplasma divulgatum* S5	96	0	76
15.2	F*erroplasma acidarmanus*	None	*Cuniculiplasma divulgatum* S5	99	0	79
15.6	*Picrophilus torridus* DSM 9790	None	*Cuniculiplasma divulgatum* S5	100	0	99
15.7	*Ferroplasma acidarmanus*	*Ferroplasma*	*Ferroplasma acidiphilum* strain Y	100	0	83
17	*Acidiphilium cryptum* JF-5	None	*Acidiphilium cryptum* JF-5	98	0	84
18.2	Uncultured methanogenic archaeon RCI	None	*Thermoplasmatales* archaeon BRNA1	89	9e-100	70
19.2	*Thermoplasma volcanium* GSS1	None	*Cuniculiplasma divulgatum* S5	99	0	79
21.2	*Acidimicrobium ferrooxidans* DSM 10331	None	*Acidimicrobium ferrooxidans* DSM 10331	89	3e-157	73
22.1	*Geobacter metallireducens* GS-15	*Leptospirillum*	*Leptospirillum ferrooxidans* C2-3	100	0	100
22.2	*Geobacter metallireducens* GS-15	*Leptospirillum*	*Leptospirillum ferrooxidans* C2-3	100	0	93

24	*Sulfobacillus thermosulfidooxidans*	None	*Sulfobacillus acidophilus* DSM 10332	94	1e-173	74
26	*Thermoplasma acidophilum* DSM 1728	None	Uncultured archaeon clone ASS_A1	26	1e-08	81
32	*Acidimicrobium ferrooxidans* DSM 10331	None	*Acidimicrobium ferrooxidans* DSM 10331	87	3e-138	72
34.2	*Thermoplasma acidophilum* DSM 1728	None	Uncultured archaeon clone ASS_A1	95	0	76
39	*Desulfuromonas acetoxidans*	*Leptospirillum*	*Leptospirillum ferrooxidans* C2-3	99	0	84
41	*Thermoplasma acidophilum* DSM 1728	None	Uncultured archaeon clone ASS_A1	92	2e-158	74
43	*Moorella thermoacetica*	None	*Thermus thermophilus*	35	9e-100	77
44	*Thermococcus kodakarensis* KOD1	None	Candidatus *Micrarchaeota* archaeon Mia14	99	1e-97	69
49	*Archaeoglobus fulgidus* DSM 4304	None	Candidatus *Micrarchaeota* archaeon Mia14	70	2e-77	70
52	*Thioalkalivibrio* sp. HL-EbGR7	None	*Thioalkalivibrio sulfidiphilus* HL-EbGr7	94	0	82
56	*Thermoplasma volcanium* GSS1	None	*Cuniculiplasma divulgatum* strain PM4	96	0	77
57	*Ferroplasma acidarmanus*	None	*Cuniculiplasma divulgatum* strain PM4	99	1e-174	74

Table 5.6 Taxonomic information for genomes constructed from Wheal Maid metagenomic sequence data (with < 85 % completion).

Genome	Genus (Anvi'o)	*recA* BLAST analysis			
		Species	% query coverage	E value	% identity
2.1	*Acidithiobacillus*	*Acidithiobacillus ferrivorans* SS3	100	0	92
2.4	*Acidithiobacillus*	*Acidithiobacillus ferrivorans* SS3	100	0	99
5.1	None	*Acidimicrobium ferrooxidans* DSM	90	0	77
5.2	None	*Ilumatobacter coccineus* YM16-304	90	0	79
7	None	*Ilumatobacter coccineus* YM16-304	90	0	77
8	None	*Acidimicrobium ferrooxidans* DSM	85	0	78
9	None	*Acidobacterium capsulatum* ATCC 51196	73	3e-42	76
10	None	*Steroidobacter denitrificans* DSM	92	0	84
12.2	None	*Caldithrix abyssi* DSM	58	7e-95	73
14.2	None	*Thermaerobacter marianensis* DSM	81	1e-179	76
16	None	*Acidobacterium capsulatum* ATCC 51196	94	7e-158	81
21.1	None	*Ilumatobacter coccineus* YM16-304	87	0	77
23	*Leptospirillum*	*Leptospirillum* sp. Group II	99	0	84
25	None	*Terriglobus roseus* DSM 18391	88	8e-75	78
28	None	*Clostridium pasteurianum* M150B	83	4e-110	70
30	None	*Methanobrevibacter* sp. AbM4	54	2e-25	66
31	None	*Leptospirillum* sp. Group II 'CF-1	98	0	79
33	None	*Leptospirillum* sp. Group II 'CF-1	82	6e-172	75
34.1	None	Candidatus *Micrarchaeota* archaeon Mia14	98	4e-110	71
35	None	Candidatus *Babela massiliensis* strain BABL1	77	3e-99	71
36	None	*Agarivorans gilvus* WH0801	84	3e-68	
37	None	*Acidimicrobium ferrooxidans* DSM 10331	90	0	76
40	None	*Thermoanaerobacterium xylanolyticum* LX-1162	62	3e-74	70

5.5 Genes related to metal resistance, a key characteristic in organisms which can thrive in AMD, were found across the samples.

AMD environments are typically contaminated with high levels of metals which have been mobilised from rocks and minerals by the acidic conditions. Microorganisms which thrive in AMD must therefore have resistance to a range of metals. In this study genes implicated in resistance to mercury, arsenic, cadmium, zinc and copper were sought in annotated sequence data. Where possible these were looked for in both individual genomes extracted from the metagenomic dataset (presence or absence of genes recorded), as well as in the community as a whole at site 1, surface level and depth, and site 2 surface level (abundance of genes, based on reads per kilobase per million (rpkm) recorded).

5.5.1 Arsenic resistance

Arsenic has been identified as a major pollutant at the Wheal Maid tailings lagoon (Carrick District Council, 2008) and is often found in high levels at AMD sites globally (Johnson & Hallberg, 2005). Arsenic is toxic to most microorganisms, however genes have been identified within AMD populations which are involved in resistance to and the bioremediation of arsenic. The *ars* operon contains genes which enable the transportation of arsenic out of cells; *arsA* and *arsB* form an anion-translocating ATPase which catalyses extrusion out of the cell of the oxyanions arsenite, antimonite and arsenate (Rosen *et al*., 1990). However, arsenate must be reduced to arsenite prior to extrusion and this process is controlled by the arsenate reduction gene *arsC* (Martin *et al*., 2001). *aox* genes allow for the oxidation of arsenite to the less toxic arsenate. The operon is regulated by *arsD* and *arsR*. An additional *ars* gene, *arsH*, is also sometimes found on this operon and, until recently, its role was not understood, however it is now believed to be responsible for detoxifying trivalent methylated and aromatic arsenicals by oxidation (Chen *et al.*, 2015). *ars* genes were looked

for in gene annotation data generated by RAST and FMAP, both in individual genomes extracted from the Wheal Maid metagenome, as well as across the three samples as a whole. Across individual genomes *arsA*, *arsB*, *arsC*, *arsD*, *arsH* and *arsR* were found (Table 5.7). Across all samples *arsA*, *arsB*, *arsC*, *arsH* and *arsR* (no annotation was available for *arsD*) were also found, along with *aoxA, aoxB* and *acr3* (Table 5.8). Only one genome (Bin 17) had the *arsH* gene present, however it is present in relatively high abundance across all three samples. The only gene completely absent from a sample is *arsC*, which is not present at site 1, depth.

5.5.2 Mercury resistance

Mercury exists in only very small amounts naturally in the environment, however anthropogenic activities, including mining, have led to increased levels of mercury in certain areas (Dash & Das, 2012). Data is not available regarding mercury levels at the Wheal Maid site, however mercury is a known contaminant within AMD at other mining sites, and knowledge of microorganisms which are able to thrive in AMD and are resistant to mercury is beneficial when looking at which bacteria can be used in bioremediation processes. Mercury-resistant bacteria harbour the *mer* operon in their genome which enables bacteria to detoxify Hg^{2+} into volatile metallic mercury (Nascimento & Chartone-Souza, 2003). Genes present in the *mer* operon include : *merR or merD*, which regulates the operon and one or more of *merT*, *merC*, *merE*, *merF* and *merG*, which transport mercury to the cytoplasm for reduction by *merA*, the central enzyme in the detoxification of mercury which catalyses the reduction of Hg(II) to volatile Hg(0) (Boyd & Barkay 2012). *mer* genes present in individual genomes and across samples are shown in Tables 5.9 and 5.10.

Table 5.7 Presence or absence of arsenic resistance (*ars*) genes in genomes constructed from Wheal Maid metagenomic sequence data.

Genome (Bin)	*arsA*	*arsB*	*arsC*	*arsD*	*arsH*	*arsR*
4.2	-	+	+	-	-	+
4.3	+	+	+	-	-	+
11.2	+	+	+	-	-	+
12.1	-	+	+	-	-	+
13	-	+	+	-	-	+
14.1	+	+	+	+	-	+
15.1	-	+	+	-	-	-
15.2	+	+	+	-	-	-
15.6	+	+	+	-	-	-
17	-	+	+	-	+	-
19.2	+	+	+	-	-	-
21.2	-	+	+	-	-	+
22.1	+	+	+	+	-	+
32	+	+	+	+	-	+
39	+	+	+	+	-	+
41	+	+	+	-	-	+
43	-	+	+	-	-	-
49	+	+	+	-	-	+
52	-	-	+	-	-	+
56	-	+	+	-	-	-
57	-	+	+	-	-	+

Table 5.8 Abundance of arsenic resistance (*ars*) and arsenic oxidation (*aox*) genes (rpkm) in metagenomic sequence datasets from two sites at Wheal Maid.

Sample	*arsA* (rpkm)	*arsB* (rpkm)	*arsC* (rpkm)	*arsH* (rpkm)	*arsR* (rpkm)	*acr3* (rpkm)	*aoxA* (rpkm)	*aoxB* (rpkm)
Site 1, surface	288.51	556.20	59.05	544.27	1039.38	1874.13	254.37	243.64
Site 1, depth	504.24	539.07	0	544.20	489.85	1811.56	205.61	116.06
Site 2, surface	685.32	518.36	213.33	239.67	2791.84	274.72	165.37	66.54

Table 5.9 Presence or absence of mercury resistance (*mer*) genes in genomes constructed from Wheal Maid metagenomic sequence data.

Genome	*merA*	*merB*	*merC*	*merP*	*merR*	*merT*
3.4	+	-	-	-	+	+
4.2	+	-	-	-	-	-
4.4	-	-	-	-	-	-
11.2	+	-	-	-	-	-
12.1	+	-	-	-	-	-
13	+	-	+	+	+	+
14.1	+	+	-	-	+	-
15.1	+	-	-	-	-	-
15.2	+	-	-	-	-	-
15.6	-	-	-	-	-	-
15.7	+	-	-	-	-	-
17	+	-	-	-	+	-
18.2	+	-	-	-	-	-
19.2	+	-	-	-	-	-
21.2	-	-	-	-	-	-
22.1	+	-	-	-	+	+
22.2	+	-	-	-	-	-
24	+	-	-	-	-	-
26	+	-	-	-	-	-
30	+	-	-	-	-	-
32	+	+	-	-	-	-
34.2	+	-	-	-	-	-
39	+	-	-	-	-	-
41	+	-	-	-	-	-
43	-	-	-	-	-	-
49	-	-	-	-	-	-
52	+	-	-	-	-	-
56	+	-	-	-	-	-
57	+	-	-	-	-	-

Table 5.10 Abundance of mercury resistance (*mer*) genes (rpkm) in metagenomic sequence datasets from two sites at Wheal Maid.

Sample	*merA* (rpkm)	*merB* (rpkm)	*merC* (rpkm)	*merD* (rpkm)	*merE* (rpkm)	*merP* (rpkm)	*merR* (rpkm)	*merT* (rpkm)
Site 1, surface	702.17	0	288.55	70.56	0	90.50	55.4	69.39
Site 1, depth	802.70	0	0	0	0	362.06	63.2	0
Site 2, surface	1244.21	0	96.98	117.19	0	0	77.1	202.35

5.5.3 Zinc, cadmium, cobalt and copper resistance

Zinc, cadmium and copper are all common contaminants within AMD systems and resistance to these metals is important if a microbial population is to thrive in the AMD environment.

Copper is an essential trace element for all living organisms, however the high levels frequently found at AMD sites are highly toxic to most organisms. P-type ATPases of the P(IB)-subclass play a major role in metal homeostasis. Extremophiles used in the biomining of copper have been shown to contain two P(IB)-ATPases, CopA and CopB, which export copper from the cell; the presence of the *copA* gene is one of the core determinants for microbial Cu-resistance and along with *copB* was looked for across the Wheal Maid metagenomic sequence dataset (Table 5.12). The presence of *copA* across all samples suggests the presence of copper resistant microorganisms. *copB* was also present at site1 and site 2 surface, but absent from site 1 depth.

Resistance to cobalt, zinc and cadmium in microorganisms is linked; the *czc* system is composed of genes encoding a multi-protein complex associated with a high level resistance to cadmium, cobalt and zinc in bacteria. The *czc* system is regulated by the cobalt-zinc-cadmium efflux system protein CzcD which regulates transmembrane proteins CzcA, CzcB and CzcC, responsible for transporting metals from the cell (Anton *et al*., 1999; Intorne *et al*., 2012). Tables 5.11 and 5.12 show that the *czc* system is present across the Wheal Maid samples and was also detected in individual genomes.

Table 5.11 Abundance of copper resistance (*cop*) genes and genes from the *czc* system (cadmium, zinc and cobalt resistance) in metagenomic sequence datasets from two sites at Wheal Maid

Sample	Copper resistance		Cadmium, cobalt zinc resistance (*czc* system)			
	copA (rpkm)	*copB* (rpkm)	*czcD* (rpkm)	*czcA* (rpkm)	*czcB* (rpkm)	*czcC* (rpkm)
Site 1, surface	474.76	112.67	392.83	559.6	66.70	58.32
Site 1, depth	459.78	0	438.43	1022.48	147.89	156.9
Site 2, surface	428.75	115.40	438.40	448.61	73.81	0

Table 5.12 Presence or absence of genes from the *czc* system (cadmium, zinc and cobalt resistance) in genomes constructed from Wheal Maid metagenomic sequence data

Genome (Bin)	*czcD*	*czcA*	*czcB*	*czcC*
3.4	+	+	+	+
4.2	+	-	+	-
4.3	+	+	+	-
4.4	+	-	+	+
11.2	+	-	-	
12.1	+	+	+	+
13	+	+	-	+
14.1	+	-	-	-
15.1	+	-	-	-
15.2	+	-	-	-
15.7	+	-	-	-
17	+	+	+	+
18.2	+	-		-
19.2	+	-	-	-
21.2	+	-	-	-
22.1	+	+	+	+
22.2	+	-	+	+
24	+	+	+	-
28	-	-	+	+
32	+	-	-	-
39	+	+	+	+
49	+	-	-	-
52	+	+	+	+

5.6 Genes required for nitrogen and carbon fixation, crucial functions within an AMD microbial population, are present in the samples.

AMD environments are typically very low in organic nitrogen and carbon. Therefore, an essential component in the microbial ecosystem is the presence of organisms capable of nitrogen and carbon fixation (Chen *et al*., 2016). Both nitrogen and carbon fixation in these systems is likely carried out by small numbers of keystone species that are present in low abundances, including *Leptospirillum* and *Acidithiobacillus* species (Chen *et al*., 2015; Hua *et al*., 2015). The ability of a prokaryote to be able to fix nitrogen is related to the presence of nitrogen fixation (*nif*) genes. *nif* genes were only identified in two of the genomes with over 85 % completion extracted from the Wheal Maid metagenomic sequence data (Bin 39 and Bin 3.4). Within these genomes the key nifA gene, responsible for *nif* transcription initiation, was present along with *nifB*, *nifD*, *nifE, nifH*, *nifK*, *nifN*, *nifQ*, *nifT*, *nifW*, *nifX* and *nifZ*. *nif* genes were also looked for across samples grouped together by site/depth: Table 5.13 shows abundance (in rpkm) across samples.

Genes involved in the Calvin cycle, a CO_2 fixation system used by most photo- and chemo-autotrophic bacteria, were only present in one genome with over 85 % completion extracted from the Wheal Maid metagenomic dataset: genes encoding PRK, RuBisCo, PGK, GAPDH, TPI, FBA, FBP, TK, RPE, RisA and RisB were all present in Bin 24. However, pathways for carbon fixation are present across all samples. Figures 5.8 – 5.13 show two different KEGG pathways for carbon fixation in each sample: Figures 5.8, 5.10 and 5.12 show general carbon fixation pathways in prokaryotes, while Figures 5.9, 5.11 and 5.13 show carbon fixation pathways in photosynthetic organisms, including the Calvin cycle. The pathways involved in carbon fixation have various levels of

completion – the Calvin cycle being most complete – with genes present across both the bacterial and archaeal members of the population.

The absence of *nif* genes and those involved in carbon fixation in the majority of genomes constructed from the Wheal Maid metagenomic dataset indicates that a limited number of species are likely responsible for maintaining the balance of nitrogen and carbon in the Wheal Maid AMD ecosystem.

Table 5.13 Abundance of nitrogen fixation genes in metagenomic sequence datasets from two sites at Wheal Maid

Sample	*nifA* (rpkm)	*nifB* (rpkm)	*nifD* (rpkm)	*nifE (rpkm)*	*nifH* (rpkm)	*nifK* (rpkm)	*nifN* (rpkm)	*nifQ* (rpkm)	*nifT* (rpkm)	*nifU* (rpkm)	*nifV* (rpkm)	*nifW* (rpkm)	*nifX* (rpkm)	*nifZ* (rpkm)
Site 1, surface	122.79	159.58	205.53	107.44	255.59	210.14	89.62	38.37	362.03	1823.03	147.16	75.69	123.42	453.08
Site 1, depth	60.56	127.38	205.05	59.48	0	259.48	72.72	0	0	1140.38	164.90	282.28	241.37	0
Site 2, surface	1419.37	513.80	465.24	362.00	236.60	299.55	618.97	0	1019.09	1054.12	176.62	127.85	127.85	769.34

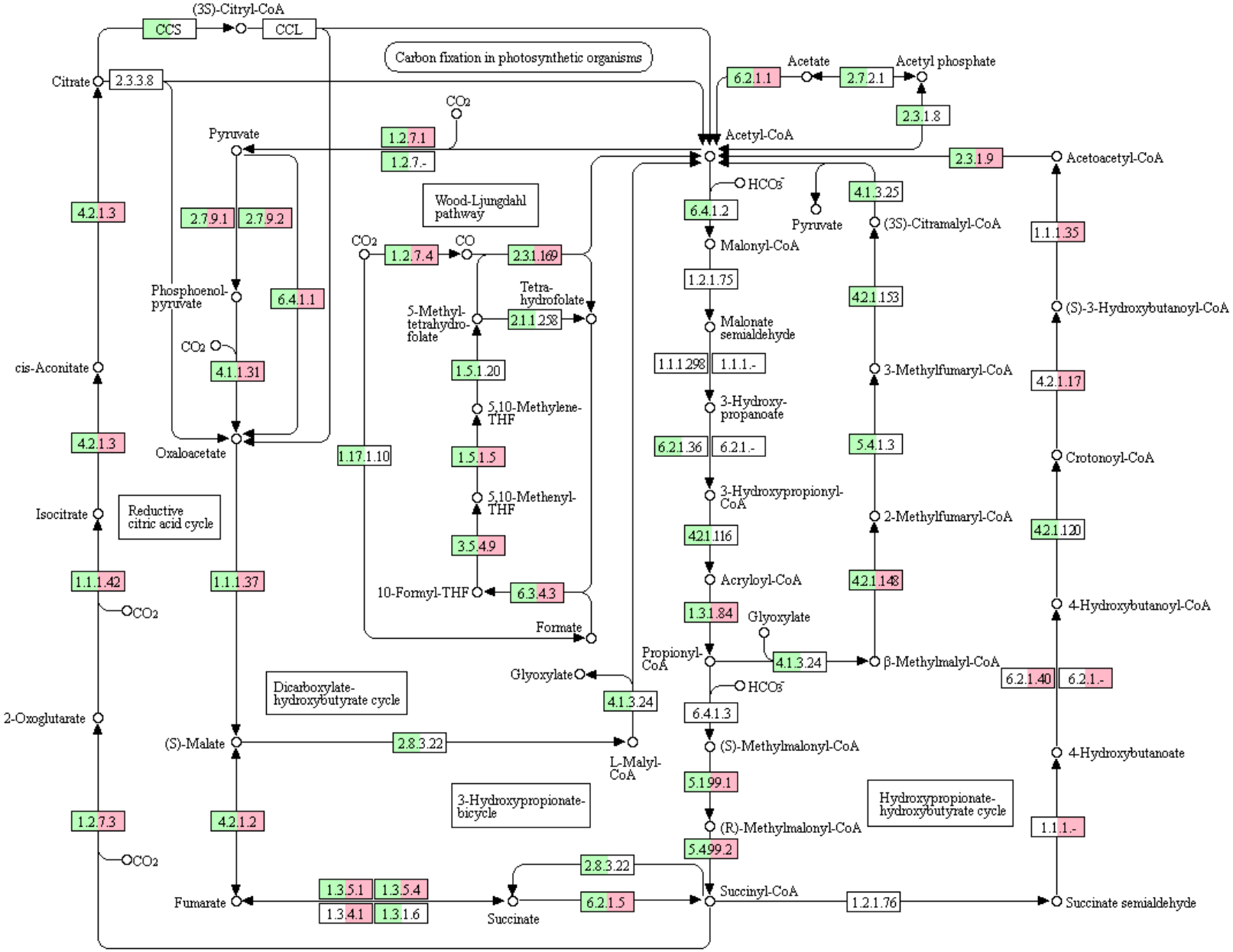


Figure 5.8 Site 1, surface level carbon fixation pathways, mapped using KEGG. Green indicates the presence of genes in bacteria and pink indicates the presence of genes in archaea from the microbial community.

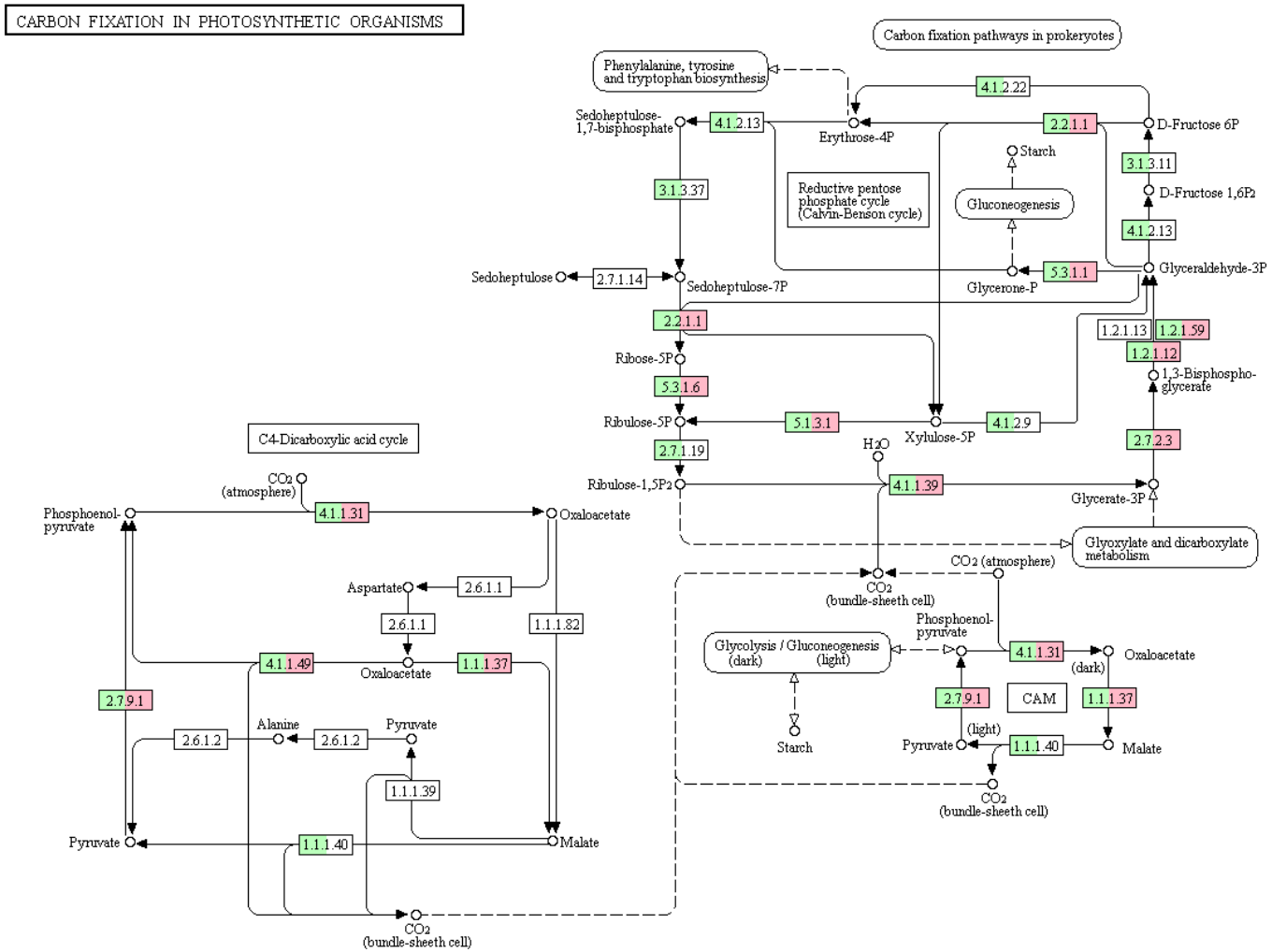


Figure 5.9 Site 1, surface level carbon fixation in photosynthetic organisms pathways, mapped using KEGG. Green indicates the presence of genes in bacteria and pink indicates the presence of genes in archaea from the microbial community.

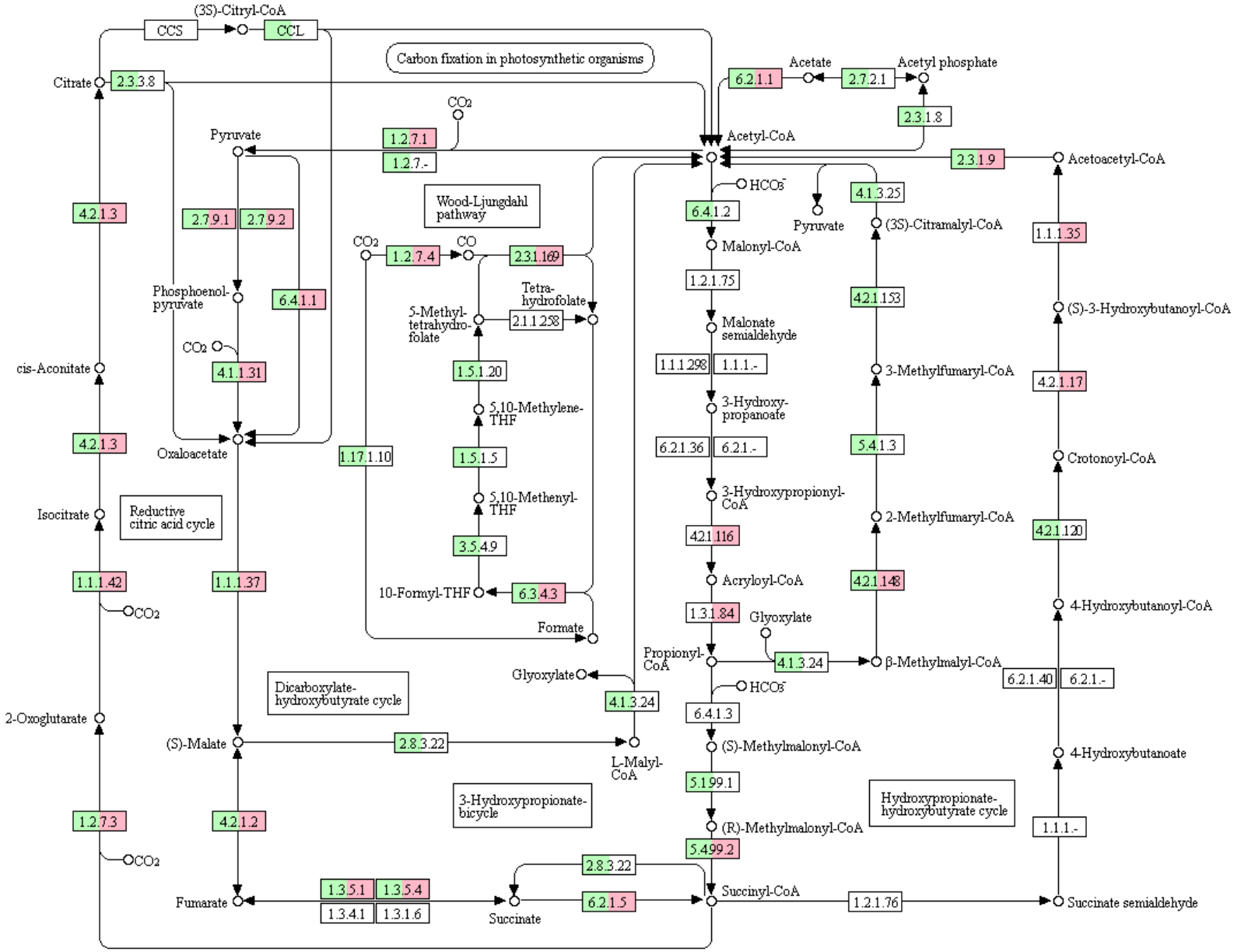


Figure 5.10 Site 1, depth carbon fixation pathways, mapped using KEGG. Green indicates the presence of genes in bacteria and pink indicates the presence of genes in archaea from the microbial community.

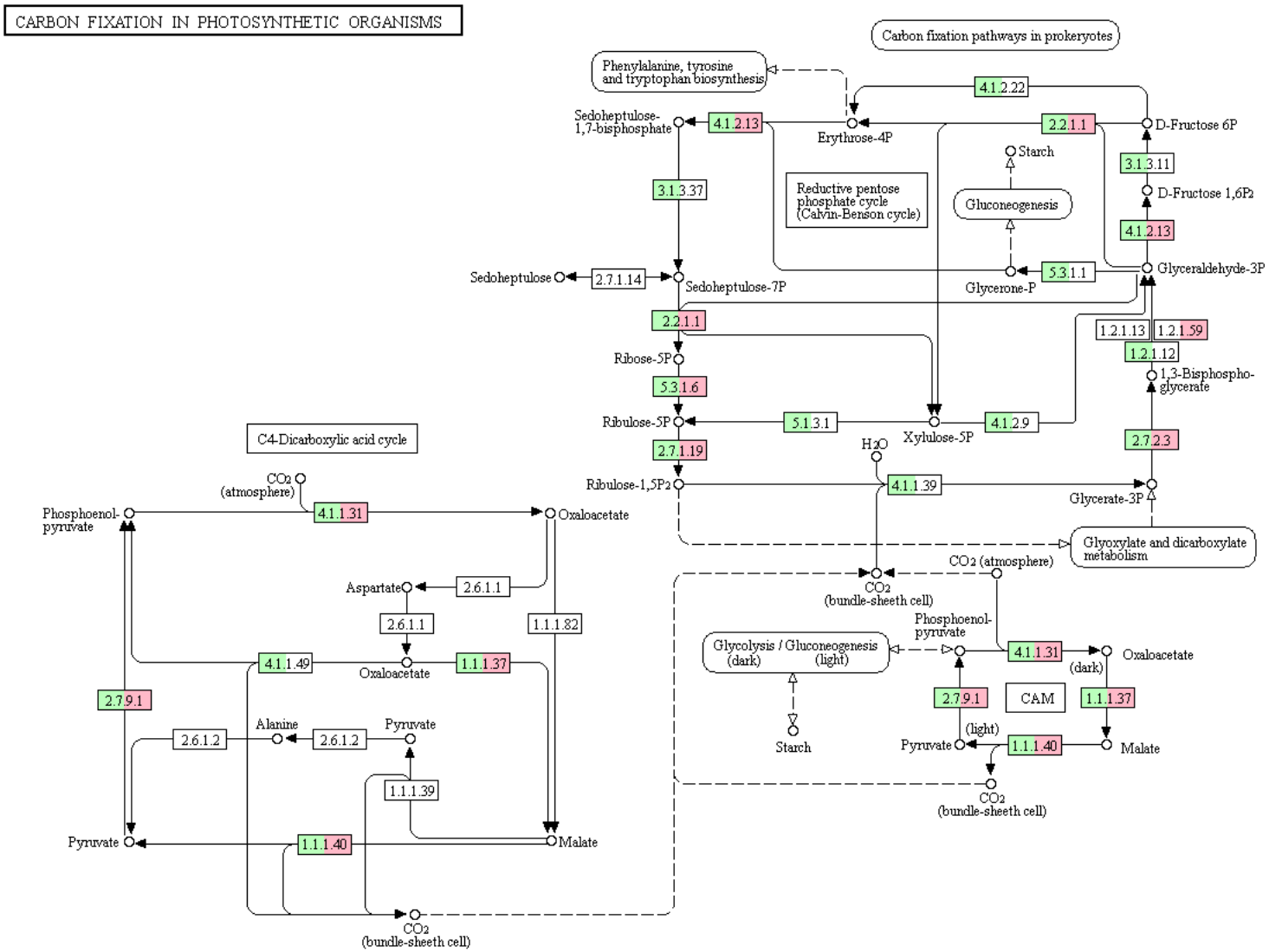


Figure 5.11 Site 1, depth carbon fixation in photosynthetic organisms pathways, mapped using KEGG. Green indicates the presence of genes in bacteria and pink indicates the presence of genes in archaea from the microbial community.

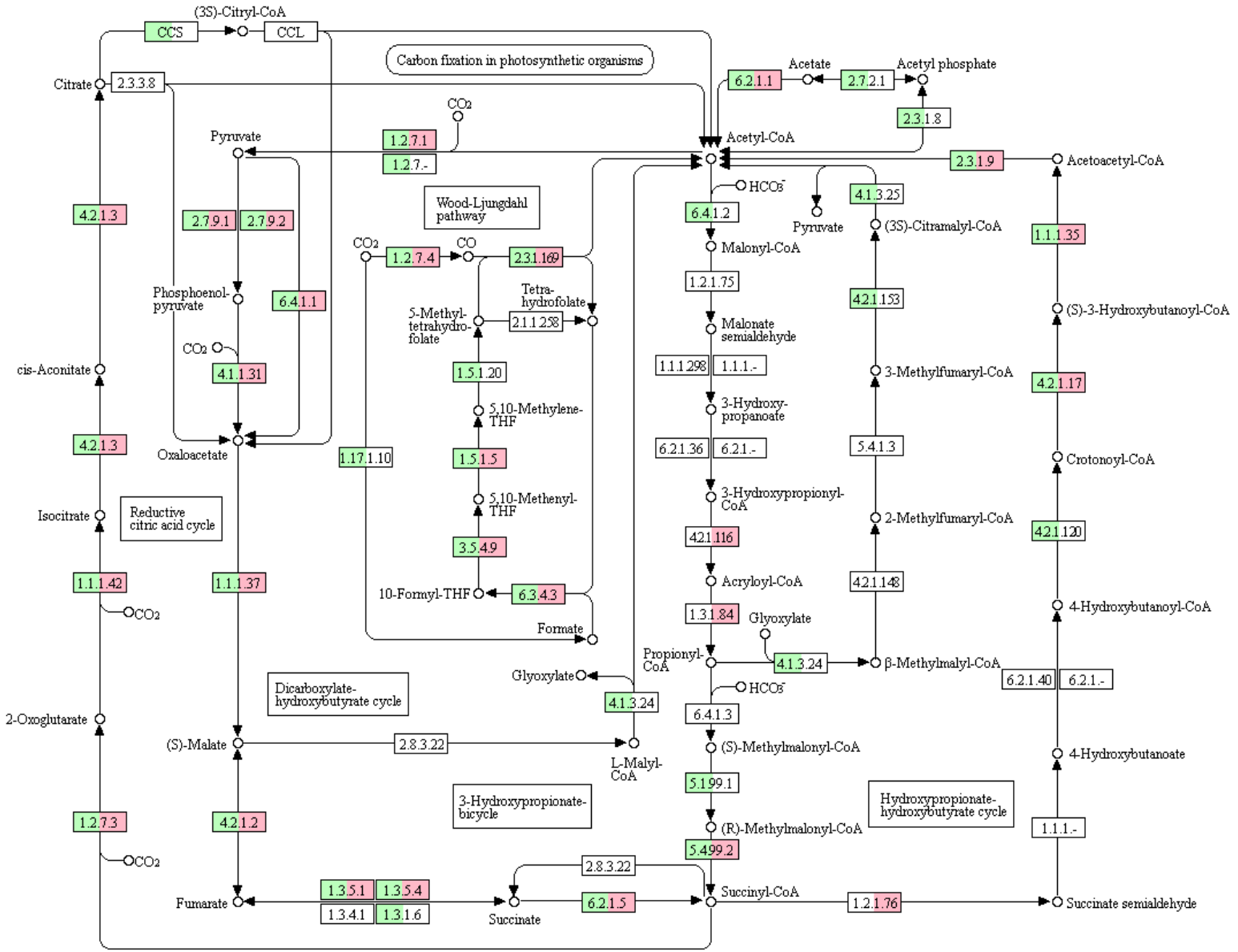


Figure 5.12 Site 2, surface level carbon fixation pathways, mapped using KEGG. Green indicates the presence of genes in bacteria and pink indicates the presence of genes in archaea from the microbial community.

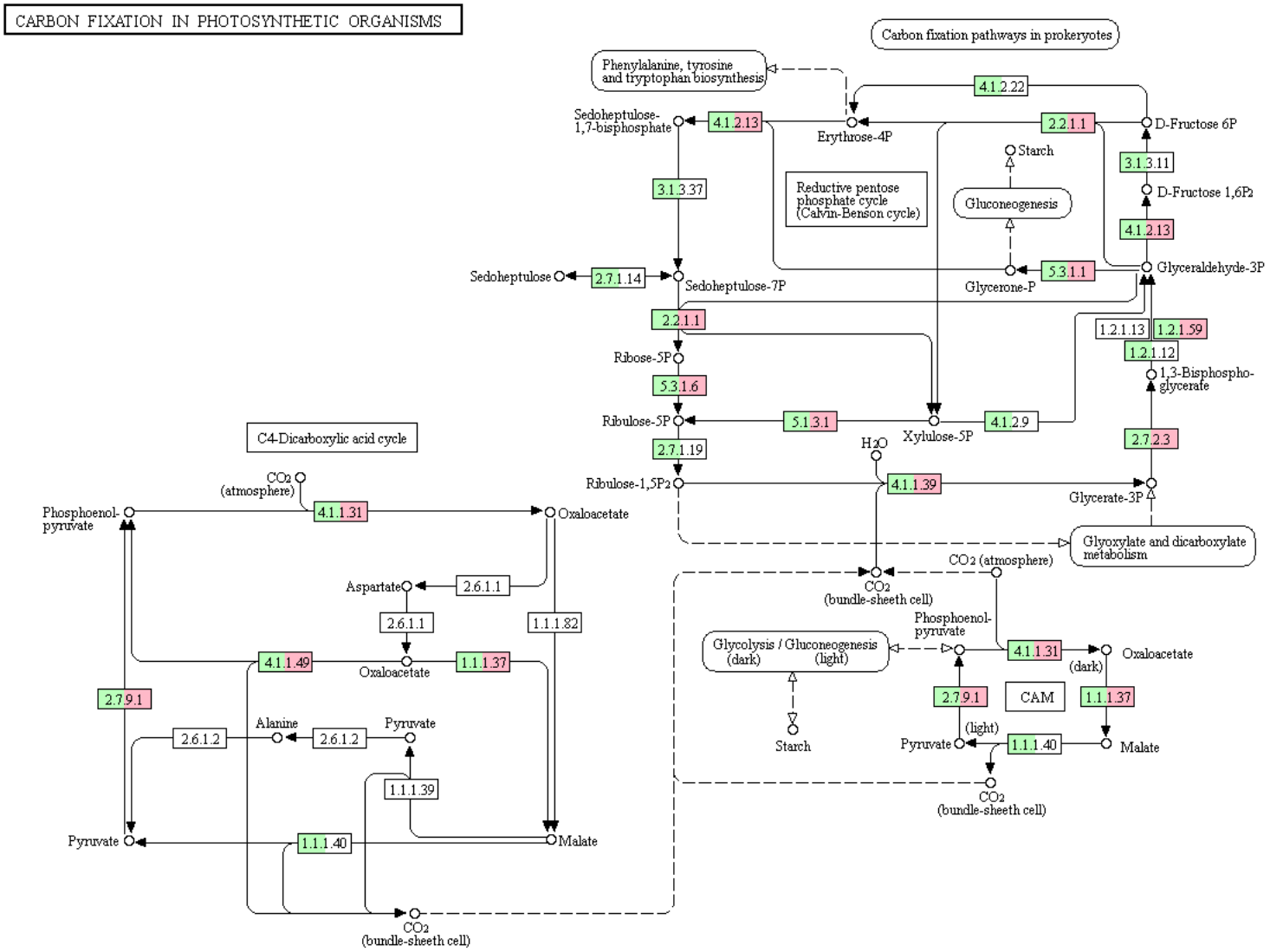


Figure 5.13 Site 2, surface level carbon fixation in photosynthetic organisms pathways, mapped using KEGG. Green indicates the presence of genes in bacteria and pink indicates the presence of genes in archaea from the microbial community.

5.7 Iron and Sulphur metabolism

Iron and Sulphur metabolism are key functions in an AMD microbial community; iron and sulphur oxidation are essential processes for chemolithoautotrophic acidophiles to obtain energy as well as key processes in the generation of AMD *(Johnson & Hallberg 2003; Johnson & Hallberg, 2008).* Additionally, the reduction of sulphate to sulphide by SRB is a process which raises the pH of AMD systems, causing metals to precipitate out of solution, and has been utilised for bioremediation purposes.

No SRB were identified from the genomes extracted from the Wheal Maid metagenomic dataset (Section 5.4). However, genes present across all three samples (Wheal Maid site 1, surface; Wheal Maid site 1, depth and Wheal Maid site 2, surface) were mapped against the KEGG pathway for sulphur metabolism (Figures 5.14-5.16; complete pathways are present for assimilatory and dissimilatory sulphate reduction. The dissimilatory sulphate reduction pathway is present in both the bacterial and the archaeal members of the microbial community across the two sites and depths. The assimilatory reduction pathway is present in both bacteria and archaea at site 1 surface but small numbers of genes are missing from the archaeal members at site 1, depth and site 2, surface. The complete sulphur-oxidising Sox enzyme system is also present across all three samples; this process appears to be solely carried out by the bacterial members of the community.

Iron oxidation in acidophiles is best understood in the previously discussed model organism *Acidithiobacillus ferrooxidans,* and a range of redox proteins present in *Acidithiobacillus* spp. have been demonstrated or proposed to be involved in Fe(II) oxidation in acidophiles; these include rusticyanin, involved in electron-transfer, and this was detected across two of the Wheal Maid samples: Wheal Maid site 1, surface (88.11 rpkm) and Wheal Maid site 2, surface (1419.36 rpkm). However, other redox proteins including Cyc1 and Cyc2 were not detected.

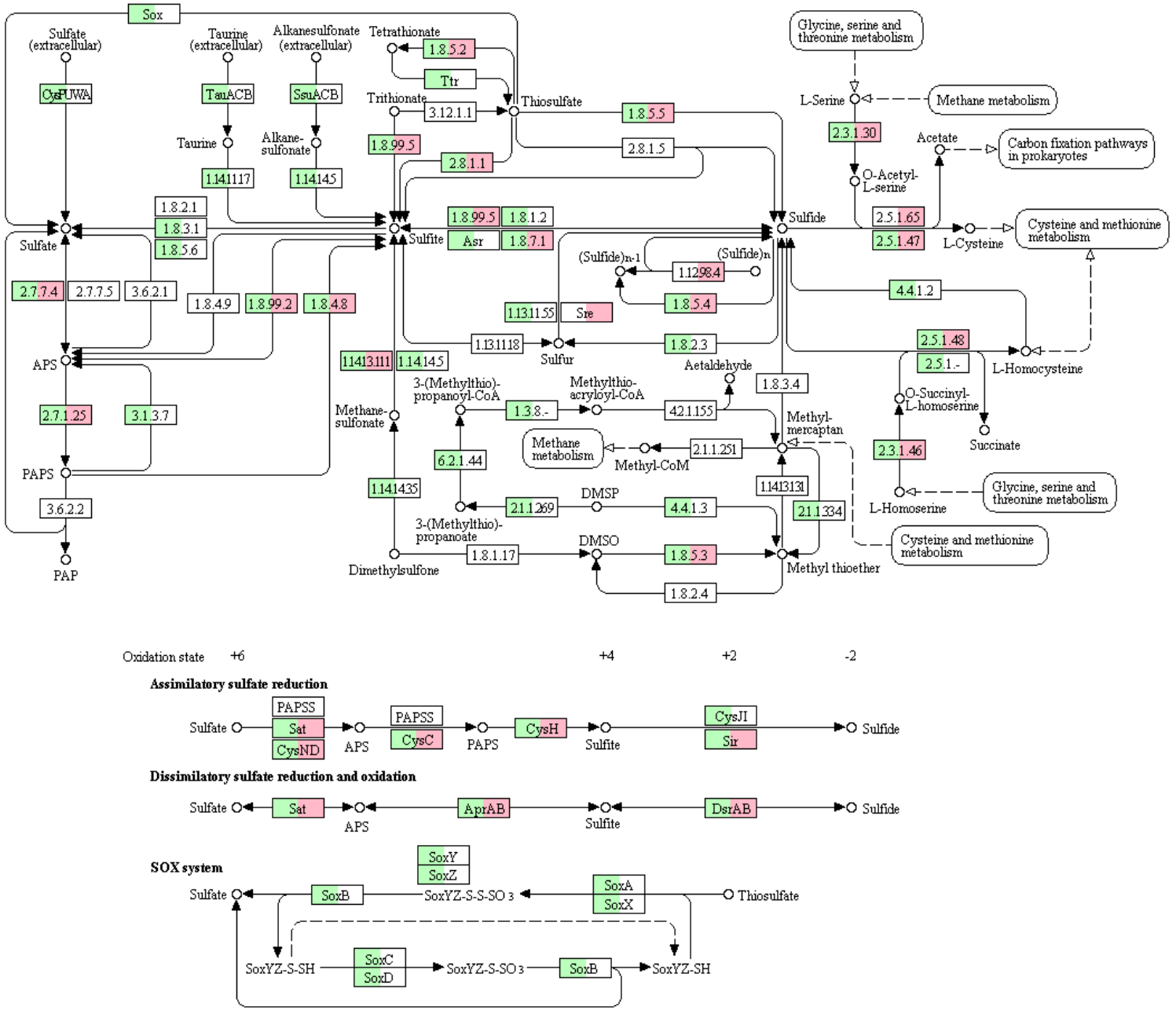


Figure 5.14 Site 1, surface level sulphur metabolism pathway, mapped using KEGG. Green indicates the presence of genes in bacteria and pink indicates the presence of genes in archaea from the microbial community.

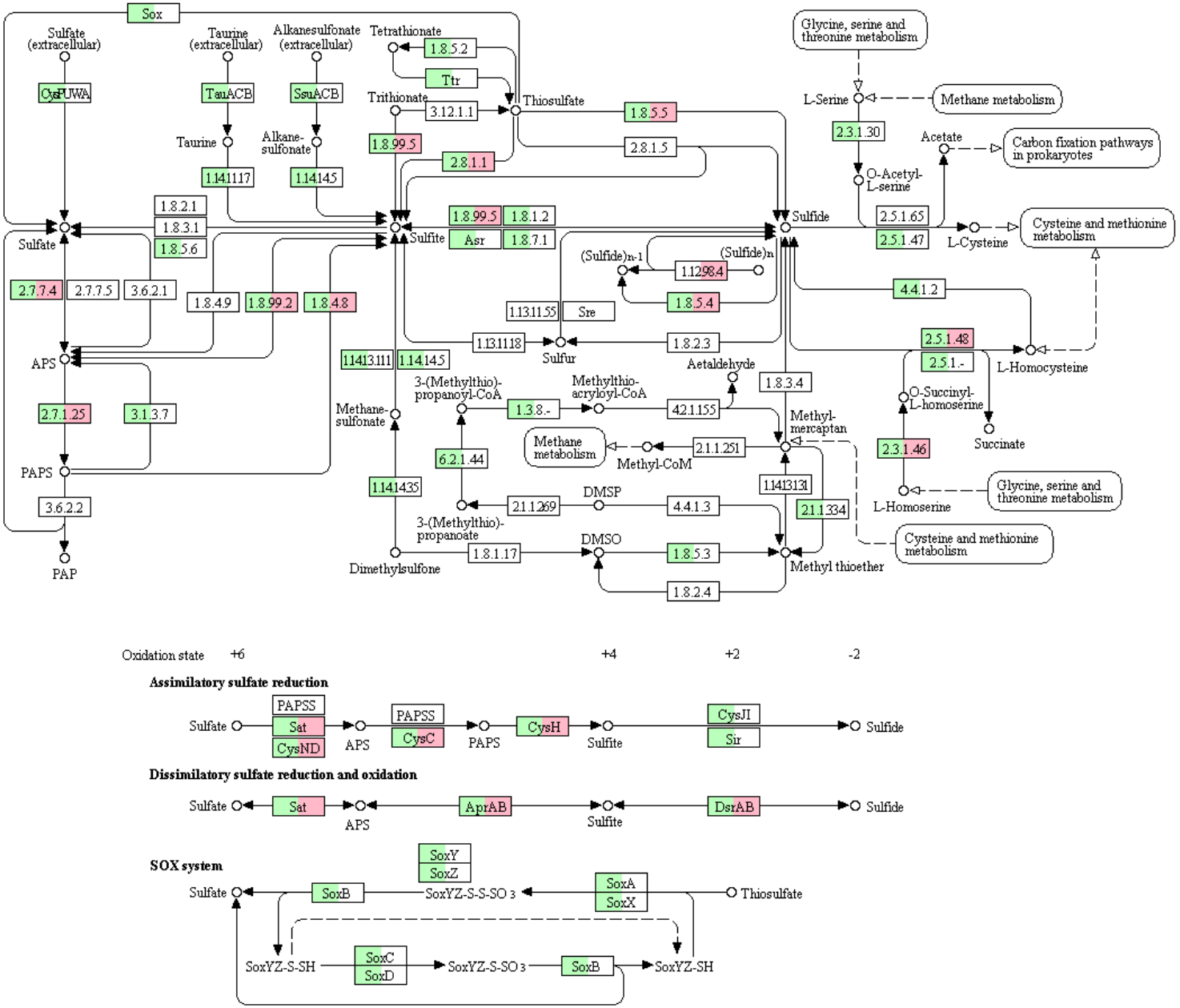


Figure 5.15 Site 1, depth sulphur metabolism pathway, mapped using KEGG. Green indicates the presence of genes in bacteria and pink indicates the presence of genes in archaea from the microbial community

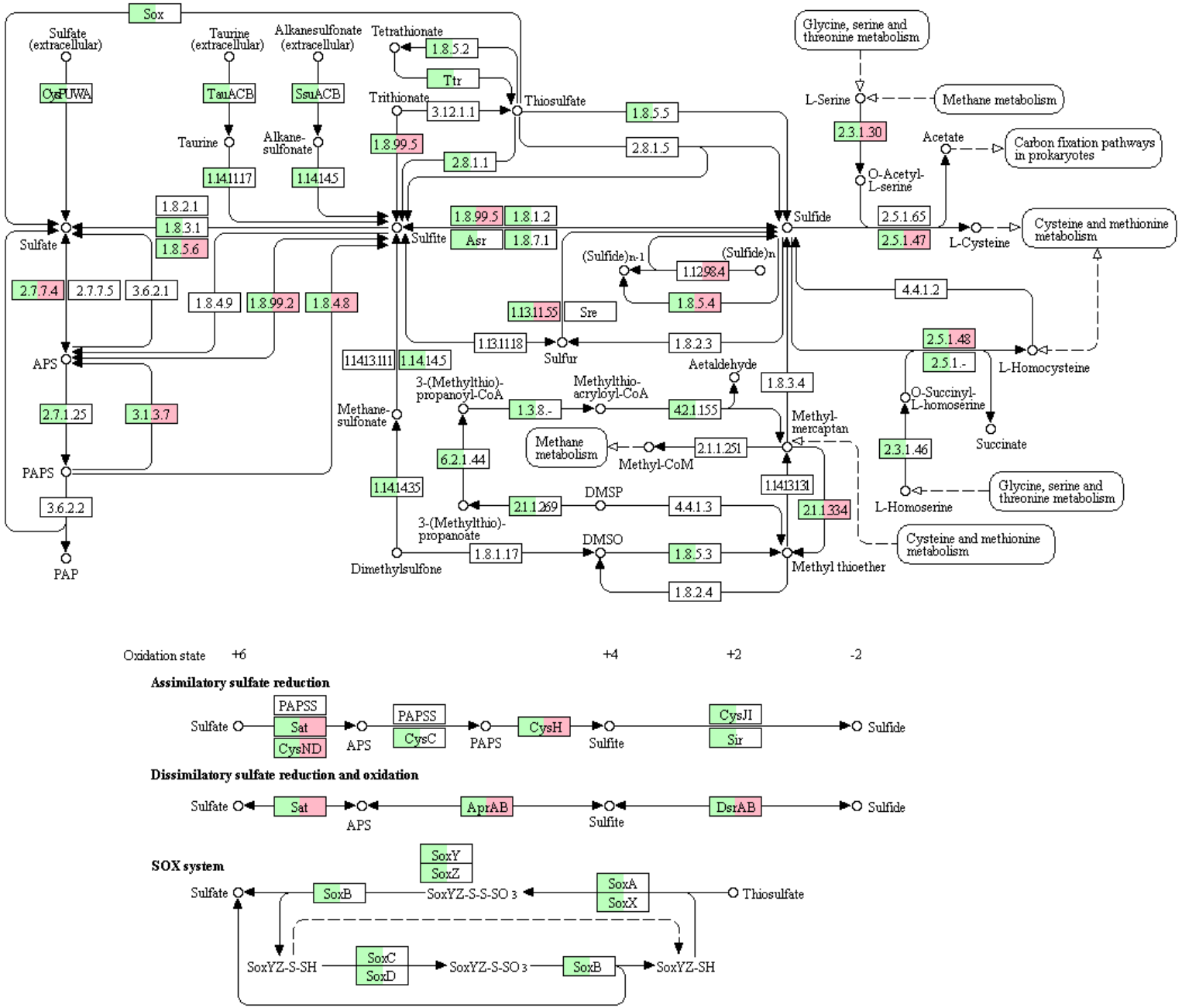


Figure 5.16 Site 2, surface level sulphur metabolism pathway, mapped using KEGG. Green indicates the presence of genes in bacteria and pink indicates the presence of genes in archaea present at the site.

5.8 Summary

Shotgun metagenomics were used to study the microbial community at two locations at the Wheal Maid AMD tailings lagoon. Samples were taken from two depths at each site, however it was not possible to create sequencing libraries from site 2 at depth. This could be due to the presence of substances at this location which inhibited the chemicals used in DNA extraction and library preparation processes. This highlights the difficulties associated with studying extreme, unknown environments. Furthermore, high numbers of sequencing reads could not be classified, and reasons for this are not fully understood; a proportion of reads not classified appeared to be novel archaea, but many reads remain unclassified and this requires further investigation.

The microbial community in Wheal Maid sediment is not as simple as might be expected for such a niche environment. Numerous acidophilic bacteria and archaea have been detected, such as *Leptospirillum, Acidiphilium, Acidithibacilllus* and *Ferroplasma* but so have organisms not previously associated with AMD, including metabolically interesting bacteria such as *Rhodopseudomonas palustris* and *Rhodanobacter denitrificans* which are likely to play a role in maintain the balance of the ecosystem with the provision of nitrogen and carbon fixation. ARMAN-like archaea are also present and further investigations into this novel group would be recommended as part of any future studies of this site. Genes and pathways related to metal resistance, nitrogen fixation and iron and sulphur oxidation are present across the samples, and in some cases it has been possible to map these pathways in individual genomes.

Chapter six: Conclusions

The work described in this thesis has used next generation sequencing methods and a range of bioinformatic tools to investigate two different microbial communities: the microbial community living with *B. braunii* and the microbial community living in AMD at two sites in Cornwall, UK.

Within this thesis three different approaches have been used to analyse members of the two microbial populations: whole genome sequencing, 16S rDNA sequencing and shotgun metagenomic sequencing. Each of these has been demonstrated to have a range of benefits and drawbacks. Whole genome sequencing is a highly valuable tool in the construction of high-quality whole genomes, from which a huge amount of information regarding the organism targeted can be inferred. However, although whole genomes can be extracted from metagenomic data, the gold standard for constructing whole genomes with accuracy and good levels of coverage requires the cultivation of individual organisms. Within this thesis cultivation and whole genome sequencing was carried out on five members of the microbial population found with *B. braunii*. This enabled a large range of analysis to be carried out on these five bacteria, which were published in a genome announcement, greatly contributing to knowledge of species found alongside *B. braunii.* However, as subsequent 16S rDNA sequencing showed, the bacterial population with *B. braunii* is diverse and by only focussing on bacteria which can be cultivated, a great deal of information regarding the bacterial ecosystem is lost. With the vast majority of microorganisms in microbial communities being uncultivable, the use of cultivation and whole genome sequencing for the study of microbial populations should be used alongside other methods such as 16S rDNA or metagenomic sequencing to fully understand the ecology of a population. 16S rDNA sequencing of a community is relatively cheap (compared to metagenomic sequencing) and allows for estimations to be made regarding the taxonomic distribution of a microbial population. There are numerous tools and databases dedicated to 16S rDNA sequence data, however, the choice of method for

classification can have a significant effect on the accuracy of results and careful consideration should be given to the selection of classifiers. Additionally, differences in the variable region of the 16S rRNA gene chosen for amplification and sequencing can have an effect on the levels of taxonomic classification that can be achieved. Despite being useful for identifying the complexity of a microbial population, 16S rDNA sequencing offers no insights into metabolic function; in order to address this, the use of shotgun metagenomics can be used. Shotgun metagenomics enables both taxonomic classification and functional annotation of a microbial community, resulting in a more thorough understanding of the population than previously mentioned methods. Within this thesis metagenomics was applied to the microbial population found in AMD at Wheal Maid, enabling a greater understanding of the taxonomy and function of the community than could be achieved through 16S rDNA sequencing alone. However, despite the huge amount of information that it can provide, metagenomics is not without its drawbacks; it can be costly to achieve the depth of coverage required in order to gain useful sequence information for all members of a population. Low coverage can result in difficulties assembling and constructing whole genomes, which is a key aim when carrying out metagenomics studies.

The oil-producing bacteria B. braunii has long been of interest to the biofuel industry due its ability to produce high levels of oils in the form of hydrocarbons. However, its cultivation on an industrial scale has not, as yet, been possible due to (amongst other issues) its growth rates. *B. braunii* is known to grow with a variety or microorganisms, however studies into the identification of this microbial community and the effects it may have on the growth of B. braunii are limited. This thesis used whole genome sequencing and 16S rDNA sequencing to further knowledge in this area. Cultivation and whole genome sequencing determined the presence of novel *Shinella* spp. along with strains of *Achromobacter piechaudii*, *Agrobacterium* sp. and *Microbacterium* sp. in the *B. braunii* microbial community. These were subsequently published in a genome announcement and sequence data is available from the NCBI database (Jones *et al*., 2016). Annotation of these genomes demonstrated ways in which they may be interacting with *B. braunii* including the synthesis of B-vitamins and the presence of secretion systems which are involved in bacterial symbiosis with

plants and algae. 16S rDNA sequencing of the microbial community revealed a diverse range of bacteria live with *B. braunii*, with differences observed between those living in close association and those living in loose association with the alga. Bacteria from phyla including Acidobacteria, Actinobacteria, Bacteroidetes, Planctomycetes Proteobacteria and Verrucomicrobia were identified. From within these phyla bacteria were identified which may be in symbiotic relationships with *B. braunii,* including *Flavobacterium*, *Prosthecobacter*and and members of the order Phycisphaerales as well as bacteria which have not been previously documented as living alongside algae or whose interactions with algae are not fully understood, including *Mycobacterium* and *Rhodococcus*, both of which may be utilising the hydrocarbons produced by *B. braunii.*

AMD is a worldwide pollutant and methods to decontaminate AMD sites are continually being developed. The use of microorganisms for the bioremediation of AMD sites has shown great promise and therefore studies investigating the microbial communities found in AMD are of value. This thesis used 16S rDNA sequencing and shotgun metagenomics to gain an understanding of the microbial population found in samples taken from AMD-contaminated sediment at the Wheal Jane tailings lagoon, Cornwall. As well as organisms typical to previously studied AMD sites, such as *Leptospirillum, Acidiphilium, Acidithibacilllus* and *Ferroplasma* a number of organisms not previously documented as living in AMD were also detected, including *Conexibacter woesei, Rhodopseudomonas pulastris, Frankia* spp. and *Rhodanobacter denitrificans* which may be contributing to the ecology of the AMD community through the provision of nitrogen and/or carbon fixation. Potentially novel ARMAN-like archaea were detected at Wheal Maid; ARMAN are recently discovered archaea whose distribution is not yet fully understood and this is the first documentation of their discovery in Cornish mine sites. The microbial population in AMD at Wheal Maid has genes and pathways present which are related to metal resistance, nitrogen fixation, carbon fixation, sulphur metabolism and Iron oxidation.

This thesis has offered unique insights into the two communities studied, however there are still questions to be answered and there is room for further

work involving these communities to be carried out. Work in this thesis has focused on genomics, however there are other 'omics which could also be applied to these microbial communities. Genes of interest have been identified from the two datasets, however, it is not possible to determine which genes are actively expressed; metatranscriptomics would allow for this to be explored. Using metatranscriptomics, differences in gene expression within the bacterial community could be monitored when under different conditions; this could include altering nutrient levels within the medium of the *B. braunii* culture and looking at the effects of pH levels and toxic metal levels on the gene expression of the microbial community in AMD.

The AMD microbial community at Wheal Maid appears to host novel organisms, many of which are likely to be archaea, and investigation into this part of the community would be recommended as part of any future study of the Wheal Maid microbial community. This could be approached in numerous possible ways including: using archaea-specific primers in 16S rDNA sequencing, carrying out metagenomic sequencing with greater depth of coverage to enable more accurate genome assemblies or using long-read sequencing technologies, such as the Oxford Nanopore MinIon, to reconstruct the small genomes of ARMAN-type archaea. Additionally laboratory work in this area could focus on methods for culturing previously uncultivable organisms through co-culturing with possible host organisms or using specialised media mimicking the natural AMD environment.

Identification of bacteria which may benefit the growth and flocculation of *B. braunii* has been carried out in this thesis through whole-genome analyisis and 16S rDNA sequencing. Future work could involve creating synthetic communities composed of beneficial bacteria with which *B. braunii* could be inoculated with. Full shotgun metagenomics of the *B. braunii* community would be recommended; as well as furthering our understanding of which bacteria are present and which may be beneficial to *B. braunii* this would also assist in efforts to sequence the genome of *B. braunii* itself, which has proven difficult due to contamination from numerous bacteria.

References

Aaronson, S., Berner, T., Gold, K., Kushner, L., Patni, N. J., Repak, A., & Rubin, D. (1983). Some observations on the green planktonic alga, *Botryococcus braunii* and its bloom form. *Journal of Plankton Research*, 5, 693-700.

Ahmed, R. A., He, M., Aftab, R. A., Zheng, S., Nagi, M., Bakri, R., & Wang, C. (2017). Bioenergy application of *Dunaliella salina* SA 134 grown at various salinity levels for lipid production. *Scientific Reports*, 7, 8118.

Akcil, A., & Koldas, S. (2006). Acid Mine Drainage (AMD): causes, treatment and case studies. *Journal of Cleaner Production*, 14, 1139-1145.

Alam, M. A., Vandamme, D., Chun, W., Zhao, X., Foubert, I., Wang, Z., ... & Yuan, Z. (2016). Bioflocculation as an innovative harvesting strategy for microalgae. *Reviews in Environmental Science and Bio/Technology*, 15, 573-583.

Alikhan, N.F., Petty, N.K., Zakour, N.L., & Beatson, S.A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons, *BMC Genomics*, 12, 402.

Amaral-Zettler, L. A., Zettler, E. R., Theroux, S. M., Palacios, C., Aguilera, A., & Amils, R. (2011). Microbial community structure across the tree of life in the extreme Rio Tinto. *The ISME journal*, 5, 42.

An, D. S., Im, W. T., Yang, H. C., & Lee, S. T. (2006). *Shinella granuli* gen. nov., sp. nov., and proposal of the reclassification of *Zoogloea ramigera* ATCC 19623 as *Shinella zoogloeoides* sp. nov. *International journal of Systematic and Evolutionary Microbiology*, 56, 443-448.

Anton, A., Große, C., Reißmann, J., Pribyl, T., & Nies, D. H. (1999). CzcD is a heavy metal ion transporter involved in regulation of heavy metal resistance in *Ralstonia* sp. strain CH34. *Journal of Bacteriology*, 181, 6876-6881.

Aronesty, E. (2013). Comparison of sequencing utility programs. *The Open Bioinformatics journal*, *7*, 1-8.

Ashokkumar, V., & Rengasamy, R. (2012). Mass culture of *Botryococcus braunii* Kutz. under open raceway pond for biofuel production. *Bioresource Technology*, 104, 394-399.

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., & Weightman, A. J. (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology*, 71, 7724-7736.

Asta, M. P., Cama, J., Martínez, M., & Giménez, J. (2009). Arsenic removal by goethite and jarosite in acidic conditions and its environmental implications. *Journal of Hazardous Materials*, 171, 965-972.

Asveld, L. (2016). The need for governance by experimentation: The case of biofuels. *Science and Engineering Ethics*, 22, 815-830.

Auld, R. R., Myre, M., Mykytczuk, N. C., Leduc, L. G., & Merritt, T. J. (2013). Characterization of the microbial acid mine drainage microbial community using culturing and direct sequencing techniques. *Journal of Microbiological Methods*, 93, 108-115.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., ... & Meyer, F. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, 9, 75.

Bai, Y., Müller, D. B., Srinivas, G., Garrido-Oter, R., Potthoff, E., Rott, M., ... & Hüttel, B. (2015). Functional overlap of the *Arabidopsis* leaf and root microbiota. *Nature*. 528, 364-383.

Baker, B. J., & Banfield, J. F. (2003). Microbial communities in acid mine drainage. *FEMS Microbiology Ecology*, 44, 139-152.

Baker, B. J., Hugenholtz, P., Dawson, S. C., & Banfield, J. F. (2003). Extremely acidophilic protists from acid mine drainage host Rickettsiales-lineage endosymbionts that have intervening sequences in their 16S rRNA genes. *Applied and Environmental Microbiology*, 69, 5512-5518.

Baker, B. J., Tyson, G. W., Webb, R. I., Flanagan, J., Hugenholtz, P., Allen, E. E., & Banfield, J. F. (2006). Lineages of acidophilic archaea revealed by community genomic analysis. *Science*, 314, 1933-1935.

Baker, B. J., Comolli, L. R., Dick, G. J., Hauser, L. J., Hyatt, D., Dill, B. D., ... & Banfield, J. F. (2010). Enigmatic, ultrasmall, uncultivated Archaea. *Proceedings of the National Academy of Sciences*, 107, 8806-8811.

Banerjee, A., Sharma, R., Chisti, Y., & Banerjee, U. C. (2002). *Botryococcus braunii:* a renewable source of hydrocarbons and other chemicals. *Critical Reviews in Biotechnology*, 22, 245-279.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A., Dvorkin, M., Kulikov, A. S., Lesin, V., Nikolenko, S., Pham, s., Prjibelski, A., Pyshkin, A., Sirotkin, A., Vyahhi, N., Tesler, G., Alekseyev, M., & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19*,* 455-477.

Barley, R. W., Hutton, C., Brown, M. M. E., Cusworth, J. E., & Hamilton, T. J. (2005). Trends in biomass and metal sequestration associated with reeds and algae at Wheal Jane Biorem pilot passive treatment plant. *Science of the Total Environment*, *345*, 279-286.

Basson, A., Flemming, L. A., & Chenia, H. Y. (2008). Evaluation of adherence, hydrophobicity, aggregation, and biofilm development of *Flavobacterium johnsoniae*-like isolates. *Microbial Ecology*, 55*,* 1-14.

Behera, S., Singh, R., Arora, R., Sharma, N. K., Shukla, M., & Kumar, S. (2015). Scope of algae as third generation biofuels. *Frontiers in Bioengineering and Biotechnology*, *2*, 90.

Belcher, J. H. (1968) Notes on the physiology of *Botryococcus braunii* Kützing. *Archives of Microbiology* 61, 335-346.

Bhatia, S. K., Kim, S. H., Yoon, J. J., & Yang, Y. H. (2017). Current status and strategies for second generation biofuel production using microbial systems. *Energy Conversion and Management*, 148, 1142-1156.

Bingle, L. E., Bailey, C. M., & Pallen, M. J. (2008). Type VI secretion: a beginner's guide. *Current Opinion in Microbiology*, 11, 3-8.

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 27, 578-579.

Boetzer, M., & Pirovano, W. (2012). Toward almost closed genomes with GapFiller. *Genome Biology*, 13, R56.

Bokulich, N. A., Rideout, J. R., Mercurio, W. G., Shiffer, A., Wolfe, B., Maurice, C. F., ... & Caporaso, J. G. (2016). mockrobiota: a public resource for microbiome bioinformatics benchmarking. *mSystems*, 5, e00062-16.

Boyd, E., & Barkay, T. (2012). The mercury resistance operon: from an origin in a geothermal environment to an efficient detoxification machine. *Frontiers in Microbiology*, 3, 349.

Bratbak, G., & Thingstad, T. F. (1985). Phytoplankton-bacteria interactions: an apparent paradox? Analysis of a model system with both competition and commensalism. *Marine Ecology Progress Series*, 23-30.

Breitwieser, F. P., & Salzberg, S. L. (2016). Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. *BioRxiv*, 084715.

Brown, M., Barley, B., & Wood, H. (2007). Minewater Treatment-Technology, Application and Policy. *Water Intelligence Online*, 6, DOI:9781780402185.

Bruneel, O., Personné, J. C., Casiot, C., Leblanc, M., Elbaz-Poulichet, F., Mahler, B. J., ... & Grimont, P. A. D. (2003). Mediation of arsenic oxidation by Thiomonas sp. in acid-mine drainage (Carnoulès, France). *Journal of Applied Microbiology*, 95, 492-499.

Bubnoff, A. (2008) Next-generation sequencing: the race is on. *Cell*, 132, 721–723.

Cabrera, G., Pérez, R., Gomez, J. M., Abalos, A., & Cantero, D. (2006). Toxic effects of dissolved heavy metals on *Desulfovibrio vulgaris* and *Desulfovibrio* sp. strains. *Journal of Hazardous Materials*, 135, 40-46.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods*.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10, 421.

Camm, G. S., Butcher, A. R., Pirrie, D., Hughes, P. K., & Glass, H. J. (2003). Secondary mineral phases associated with a historic arsenic calciner identified using automated scanning electron microscopy; a pilot study from Cornwall, UK. *Minerals Engineering*, 16, 1269-1277.

Camm, G. S., Glass, H. J., Bryce, D. W., & Butcher, A. R. (2004). Characterisation of a mining-related arsenic-contaminated site, Cornwall, UK. *Journal of Geochemical Exploration*, 82, 1-15.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... & Huttley, G. A. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7, 335-336.

Carattoli A, Zankari E, Garcia-Fernandez A, Voldby Larsen M, Lund O, Villa L, Aarestrup FM , Hasman H (2014) PlasmidFinder and pMLST: in silico detection and typing of plasmids. *Antimicrobial Agents Chemother*, 58, 3895-903.

Carrasco, G., Valdezate, S., Garrido, N., Villalón, P., Medina-Pascual, M. J., & Sáez-Nieto, J. A. (2013). Identification, typing, and phylogenetic relationships of the main clinical Nocardia species in Spain according to their *gyrB* and *rpoB* genes. *Journal of Clinical Microbiology*, 51, 3602-3608.

Carrick District Council (2008) Environmental Protection Act 1990, Part2A – Section 78B Record of Determination of Wheal Maid Tailings Lagoons, Gwennap, Cornwall as Contaminated Land. Available: https://www.cornwall.gov.uk/media/3625647/2008-09-16-Record-of-Determination.pdf

Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., & Kjelleberg, S. (2007). Use of 16S *rRNA* and *rpoB* genes as molecular markers for microbial ecology studies. *Applied and Environmental Microbiology*, 73, 278-288.

Castelle, C. J., Wrighton, K. C., Thomas, B. C., Hug, L. A., Brown, C. T., Wilkins, M. J., ... & Taylor, R. C. (2015). Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Current Biology*, 25, 690-701.

Cevallos, M. A., Cervantes-Rivera, R., & Gutiérrez-Ríos, R. M. (2008). The repABC plasmid family. *Plasmid*, 60, 19-37.

Chakravorty, S., Helb, D., Burday, M., Connell, N., & Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods*, 69, 330-339.

Chatsungnoen, T., & Chisti, Y. (2016). Harvesting microalgae by flocculation–sedimentation. *Algal Research*, *13*, 271-283.

Chen, L. X., Huang, L. N., Méndez-García, C., Kuang, J. L., Hua, Z. S., Liu, J., & Shu, W. S. (2016). Microbial communities, processes and functions in acid mine drainage ecosystems. *Current Opinion in Biotechnology*, *38*, 150-158.

Chen, J., Bhattacharjee, H., & Rosen, B. P. (2015). ArsH is an organoarsenical oxidase that confers resistance to trivalent forms of the herbicide monosodium methylarsenate and the poultry growth promoter roxarsone. *Molecular Microbiology*, 96, 1042-1052.

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., ... & Cramer, G. R. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, 13, 1050.

Chirac, C., Casadevall, E., Largeau, C., & Metzger, P. (1985). Bacterial influence upon growth and hydrocarbon production of the green alga *Botryococcus braunii. Journal of Phycology*, 21, 380-387.

Chistoserdova, L. (2010) Recent progress and new challenges in metagenomics for biotechnology. *Biotechnology letters*, 32, 1351–1359

Cho, D. H., Ramanan, R., Heo, J., Lee, J., Kim, B. H., Oh, H. M., & Kim, H. S. (2015). Enhancing microalgal biomass productivity by engineering a microalgal–bacterial community. *Bioresource Technology*, 175, 578-585.

Chojnacki, S., Cowley, A., Lee, J., Foix, A., & Lopez, R. (2017). Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. *Nucleic Acids Research*, gkx273.

Clarridge, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews*, 4, 840-862.

Coburn, B., Sekirov, I., & Finlay, B. B. (2007). Type III secretion systems and disease. *Clinical Microbiology Reviews*, 20*,* 535-549.

Colodner, R., Rock, W., Chazan, B., Keller, N., Guy, N., Sakran, W., & Raz, R. (2004). Risk factors for the development of extended-spectrum beta-lactamase-producing bacteria in nonhospitalized patients. *European journal of Clinical Microbiology and Infectious Diseases*, 23, 163-167.

Compeau, P. E., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, *29*(11), 987.

Coulthurst, S. J. (2013). The Type VI secretion system–a widespread and versatile cell targeting system. *Research in Microbiology*, 164, 640-654.

Croft, M. T., Lawrence, A. D., Raux-Deery, E., Warren, M. J., & Smith, A. G. (2005). Algae acquire vitamin B 12 through a symbiotic relationship with bacteria. *Nature*, 438, 90.

Croft, M. T., Warren, M. J., & Smith, A. G. (2006). Algae need their vitamins. *Eukaryotic Cell*, 5, 1175-1183.

Dang, H., & Lovell, C. R. (2016). Microbial surface colonization and biofilm development in marine environments. *Microbiology and Molecular Biology Reviews*, 80, 91-138.

Danger, M., Leflaive, J., Oumarou, C., Ten‐Hage, L., & Lacroix, G. (2007). Control of phytoplankton–bacteria interactions by stoichiometric constraints. *Oikos*, 116, 1079-1086.

Darling, A. E., Treangen, T. J., Messeguer, X., & Perna, N. T. (2007). Analyzing patterns of microbial evolution using the mauve genome alignment system. *Comparative Genomics*, Humana press, 135-152.

Dash, H. R., & Das, S. (2012). Bioremediation of mercury and the importance of bacterial mer genes. *International Biodeterioration & Biodegradation*, 75, 207-213.

Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., & Salzberg, S. L. (1999). Alignment of whole genomes. *Nucleic Acids Research*, 11, 2369-2376.

Demirbas, A. (2010). Use of algae as biofuel sources. *Energy Conversion and Management*. 52, 16 3-170.

Demirbas, A., & Demirbas, M. F. (2011). Importance of algae oil as a source of biodiesel. *Energy Conversion and Management*, 52, 163-170.

Denison, R. F., & Kiers, E. T. (2004). Lifestyle alternatives for rhizobia: mutualism, parasitism, and forgoing symbiosis. *FEMS Microbiology Letters*, 237, 187-193.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, 72, 5069-5072.

Dimkpa, C., Weinand, T., & Asch, F. (2009). Plant–rhizobacteria interactions alleviate abiotic stress conditions. *Plant, Cell & Environment*, 32, 1682-1694.

Dini-Andreote, F., Andreote, F. D., Araújo, W. L., Trevors, J. T., & van Elsas, J. D. (2012). Bacterial genomes: habitat specificity and uncharted organisms. *Microbial ecology*, 64, 1-7.

Dopson, M., & Holmes, D. S. (2014). Metal resistance in acidophilic microorganisms and its significance for biotechnologies. *Applied Microbiology and Biotechnology*, 98, 8133-8144.

Doucette, G. J., & McGovern, E. R. Babinchak. JA 1999. Algicidal bacteria active against *Gymnodinium breve* (Dinophyceae). Bacterial isolation and characterization of killing activity. *J. Phycol*, *35*, 1447-1454.

Eloe-Fadrosh, E. A., Ivanova, N. N., Woyke, T., & Kyrpides, N. C. (2016). Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nature Microbiology*, 4, 15032.

Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3, e1319.

Escobar-Zepeda, A., Vera-Ponce de León, A., & Sanchez-Flores, A. (2015). The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in Genetics*, 6, 348.

Edwards, K. J., Bond, P. L., Gihring, T. M., & Banfield, J. F. (2000). An archaeal iron-oxidizing extreme acidophile important in acid mine drainage. *Science*, 287, 1796-1799.

Farrand, S. K., van Berkum, P. B., & Oger, P. (2003). *Agrobacterium* is a definable genus of the family Rhizobiaceae. *International journal of Systematic and Evolutionary Microbiology*, 53, 1681-1687.

Ferreira, N. L., Mathis, H., Labbé, D., Monot, F., Greer, C. W., & Fayolle-Guichard, F. (2007). n-Alkane assimilation and tert-butyl alcohol (TBA) oxidation capacity in *Mycobacterium austroafricanum* strains. *Applied Microbiology and Biotechnology*, 75, 909-919.

Figueras, M. J., Beaz-Hidalgo, R., Hossain, M. J., & Liles, M. R. (2014). Taxonomic affiliation of new genomes should be verified using average nucleotide identity and multilocus phylogenetic analysis. *Genome Announcements*, 6, e00927-14.

Fox, G. E., Wisotzkey, J. D., & Jurtshuk JR, P. (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *International Journal of Systematic and Evolutionary Microbiology*, 34, 166-170.

Fuentes, J. L., Garbayo, I., Cuaresma, M., Montero, Z., González-del-Valle, M., & Vílchez, C. (2016). Impact of microalgae-bacteria interactions on the production of algal biomass and associated compounds. *Marine Drugs*, 14, 100.

Fuerst, J. A. (2005). Intracellular compartmentation in planctomycetes. *Annu. Rev. Microbiol.*, 59, 299-328.

Fuerst, J. A., & Sagulenko, E. (2011). Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function. *Nature Reviews Microbiology*, 9*,* 403-413.

Fukunaga, Y., Kurahashi, M., Sakiyama, Y., Ohuchi, M., Yokota, A., & Harayama, S. (2009). Phycisphaera mikurensis gen. nov., sp. nov., isolated from a marine alga, and proposal of Phycisphaeraceae fam. nov., Phycisphaerales ord. nov. and Phycisphaerae classis nov. in the phylum Planctomycetes. *The Journal of General and Applied Microbiology*, 55, 267-275.

Gan, H., Lee, M., Gan, H., Halliday, N., Williams, P., Barton, H., Hudson, A. & Savka, M. (2016) Genomic potential of a new genospecies of oligotrophic *Agrobacterium* strain isolated from deep within Lechuguilla Cave, New Mexico. Conference abstract. Available: https://emam2016.sciencesconf.org/99175/document

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S.,Dopazo,J., Meyer, T. & Conesa, A. (2012). Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics*, 28, 2678-2679.

Gaunt, M. W., Turner, S. L., Rigottier-Gois, L., Lloyd-Macgilp, S. A., & Young, J. P. (2001). Phylogenies of *atpD* and *recA* support the small subunit rRNA-based classification of rhizobia. *International Journal of Systematic and Evolutionary Microbiology*, 51, 2037-2048.

Gelvin, S. B. (2003). Agrobacterium-mediated plant transformation: the biology behind the “gene-jockeying” tool. *Microbiology and Molecular Biology Reviews*, 67, 16-37.

Georgianna, D. R., & Mayfield, S. P. (2012). Exploiting diversity and synthetic biology for the production of algal biofuels. *Nature*, 488, 329-335.

Gerardo, M. L., Van Den Hende, S., Vervaeren, H., Coward, T., & Skill, S. C. (2015). Harvesting of microalgae within a biorefinery approach: a review of the developments and case studies from pilot-plants. *Algal Research*, *11*, 248-262

Gerlach, R. G., & Hensel, M. (2007). Protein secretion systems and adhesins: the molecular armory of Gram-negative pathogens. *International Journal of Medical Microbiology*, 297, 401-415.

Gilbert, J. A., and Hughes, M. (2011) Gene expression profiling: metatranscriptomics. In High-Throughput Next Generation Sequencing. *Methods in Molecular Biology*, 733, 195-205.

Giloteaux, L., Duran, R., Casiot, C., Bruneel, O., Elbaz-Poulichet, F., & Goñi-Urriza, M. (2013). Three-year survey of sulfate-reducing bacteria community structure in Carnoules acid mine drainage (France), highly contaminated by arsenic. *FEMS Microbiology Ecology*, *83*, 724-737.

Giordano, F., Aigrain, L., Quail, M. A., Coupland, P., Bonfield, J. K., Davies, R. M., ... & Yue, J. X. (2017). De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Scientific Reports*, 7, 3935.

Giovannoni, S. J., DeLong, E. F., Schmidt, T. M., & Pace, N. R. (1990). Tangential flow filtration and preliminary phylogenetic analysis of marine picoplankton. *Applied and Environmental Microbiology*, 56, 2572-2575

Gladman, S., & Seemann, T. (2009). VelvetOptimiser. *US Patent*, 2(5).

Glöckner, F. O., Yilmaz, P., Quast, C., Gerken, J., Beccati, A., Ciuprina, A., ... & Ludwig, W. (2017). 25 years of serving the community with ribosomal RNA gene reference databases and tools. *Journal of Biotechnology*, *261*, 169-176.

Goeddel, D. V., Kleid, D. G., Bolivar, F., Heyneker, H. L., Yansura, D. G., Crea, R., ... & Riggs, A. D. (1979). Expression in *Escherichia coli* of chemically synthesized genes for human insulin. *Proceedings of the National Academy of Sciences*, 76, 106-110.

Golyshina, O. V., Lünsdorf, H., Kublanov, I. V., Goldenstein, N. I., Hinrichs, K. U., & Golyshin, P. N. (2016). The novel extremely acidophilic, cell-wall-deficient archaeon *Cuniculiplasma divulgatum* gen. nov., sp. nov. represents a new family, Cuniculiplasmataceae fam. nov., of the order Thermoplasmatales. *International Journal of Systematic and Evolutionary Microbiology*, 66, 332-340.

Golyshina, O. V., Toshchakov, S. V., Makarova, K. S., Gavrilov, S. N., Korzhenkov, A. A., La Cono, V., ... & Wolf, Y. I. (2017). 'ARMAN'archaea depend on association with euryarchaeal host in culture and in situ. *Nature Communications*, 8, 60.

Gonzalez, L. E., & Bashan, Y. (2000). Increased growth of the microalga *chlorella vulgaris* when co-immobilized and co-cultured in alginate beads with the plant-growth-promoting bacterium *Azospirillum brasilense*. *Applied and Environmental Microbiology*, 66, 1527-1531.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17, 333.

Gouy, M., Guindon, S., & Gascuel, O. (2010). SeaView version 4 : a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, 27, 221-224.

Green, D. H., Echavarri-Bravo, V., Brennan, D., & Hart, M. C. (2015). Bacterial diversity associated with the coccolithophorid algae *Emiliania huxleyi* and *Coccolithus pelagicus f. braarudii, BioMed Research International*, *2015*.

Griffiths, M. J., & Harrison, S. T. (2009). Lipid productivity as a key characteristic for choosing algal species for biodiesel production. *Journal of Applied Phycology*, 21, 493-507.

Grover, J. P. (2000). Resource competition and community structure in aquatic micro-organisms: experimental studies of algae and bacteria along a gradient of organic carbon to inorganic phosphorus supply. *Journal of Plankton Research*, 22, 1591-1610.

Guazzaroni, M. E., Morgante, V., Mirete, S., & González-Pastor, J. E. (2013). Novel acid resistance genes from the metagenome of the Tinto River, an extremely acidic environment. *Environmental Microbiology*, 15, 1088-1102.

Guiry, M. D. (2012). How many species of algae are there? *Journal of Phycology*, 48, 1057-1063.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29*,* 1072-1075.

Gurung, T. B., Urabe, J., & Nakanishi, M. (1999). Regulation of the relationship between phytoplankton *Scenedesmus acutus* and heterotrophic bacteria by the balance of light and nutrients. *Aquatic Microbial Ecology*, 17, 27-35.

Hallberg, K. B., & Johnson, D. B. (2005). Microbiology of a wetland ecosystem constructed to remediate mine drainage from a heavy metal mine. *Science of the Total Environment*, *338*, 53-66.

Hallberg, K. B. (2010). New perspectives in acid mine drainage microbiology. *Hydrometallurgy*, 104, 448-453.

Hallberg, K. B., González-Toril, E., & Johnson, D. B. (2010). *Acidithiobacillus ferrivorans*, sp. nov.; facultatively anaerobic, psychrotolerant iron-, and sulfur-oxidizing acidophiles isolated from metal mine-impacted environments. *Extremophiles*, 14, 9-19.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products*. Chemistry & Biology*, 5, 245-248.

Haque, M. M., Bose, T., Dutta, A., Reddy, C. V. S. K., & Mande, S. S. (2015). CS-SCORE: Rapid identification and removal of human genome contaminants from metagenomic datasets. *Genomics*, 106, 116-121.

Harneit, K., Göksel, A., Kock, D., Klock, J. H., Gehrke, T., & Sand, W. (2006). Adhesion to metal sulfide surfaces by cells of *Acidithiobacillus ferrooxidans*, *Acidithiobacillus thiooxidans* and *Leptospirillum ferrooxidans*. *Hydrometallurgy*, 83, 245-254.

Harriott, O. T., Hosted, T. J., & Benson, D. R. (1995). Sequences of nifX, nifW, nifZ, nifB and two ORF in the Frankia nitrogen fixation gene cluster. *Gene*, 161, 63-67.

Hassanshahian, M., Ahmadinejad, M., Tebyanian, H., & Kariminik, A. (2013). Isolation and characterization of alkane degrading bacteria from petroleum reservoir waste water in Iran (Kerman and Tehran provenances). *Marine Pollution Bulletin*, 73, 300-305.

Hedlund, B. P., Gosink, J. J. & Staley, J. T. (1997). Verrucomicrobia div. nov., a new division of the bacteria containing three new species of *Prosthecobacter*. *Antonie van Leeuwenhoek,* 72, 29–38.

Hernandez, J. P., de-Bashan, L. E., Rodriguez, D. J., Rodriguez, Y., & Bashan, Y. (2009). Growth promotion of the freshwater microalga *Chlorella vulgaris* by the nitrogen-fixing, plant growth-promoting bacterium *Bacillus pumilus* from arid zone soils. *European Journal of Soil Biology*, 45, 88-93.

Herren, J. K., & Lemaitre, B. (2011). *Spiroplasma* and host immunity: activation of humoral immune responses increases endosymbiont load and susceptibility to certain Gram-negative bacterial pathogens in *Drosophila melanogaster*. *Cellular Microbiology*, 13, 1385-1396.

Hillen, L. W., Wake, L. V., & Warren, D. R. (1980). Hydrocarbon fuels from plants. *Fuel*, 59, 446-447.

Hua, Z. S., Han, Y. J., Chen, L. X., Liu, J., Hu, M., Li, S. J., ... & Shu, W. S. (2015). Ecological roles of dominant and rare prokaryotes in acid mine drainage revealed by metagenomics and metatranscriptomics. *The ISME journal*, 9, 1280.

Hugenholtz, P., Goebel, B. & Pace, N. (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology*, 180, 4765–74.

Hugenholtz, P., & Huber, T. (2003). Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *International journal of Systematic and Evolutionary Microbiology*, 53, 289-293.

Hunter, P.R., &Gaston, M.A. (1998) Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. *Journal of clinical Microbiology*, 26, 2465-6.

Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M., & Rupp, R. (2007). Dendroscop*e: An interactive viewer for large phylogenetic trees. BMC Bioinformatics*, 22, 460.

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, 17, 377-386.

Ikehara, M., Ohtsuka, E., Tokunaga, T., Taniyama, Y., Iwai, S., Kitano, K., ... & Fujiyama, K. (1984). Synthesis of a gene for human growth hormone and its

expression in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 81, 5956-5960.

Illumina (2018). Illumina Sequencing Paltforms. Available: https://emea.illumina.com/systems/sequencing-platforms.html [Accessed 16/2/18]

Illumina, inc. (2015). *Sequencing power for every scale.* [Online] Available: http://www.illumina.com/systems/sequencing.html [Accessed 7/5/15]

Intorne, A. C., de Oliveira, M. V. V., Pereira, L. D. M., & de Souza Filho, G. A. (2012). Essential role of the czc determinant for cadmium, cobalt and zinc resistance in *Gluconacetobacter diazotrophicus* PAl 5. *Int Microbiol*, 15, 69-78.

Iverson, V., Morris, R. M., Frazar, C. D., Berthiaume, C. T., Morales, R. L., & Armbrust, E. V. (2012). Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science*, 335, 587-590.

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17, 239.

Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of Clinical Microbiology*, 45, 2761-2764.

Jaspers, E. & Overmann, J. (2004). Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. *Appl Environ Microbiol*, 70, 4831–4839.

Johnson, D. B., & Hallberg, K. B. (2003). The microbiology of acidic mine waters. *Research in Microbiology*, 154, 466-473.

Johnson, D. B., & Hallberg, K. B. (2005). Acid mine drainage remediation options: a review. *Science Of the Total Environment*, 338, 3-14.

Johnston, D., Potter, H., Jones, C., Rolley, S., Watson, I., & Pritchard, J. (2008). Abandoned mines and the water environment. *Environment Agency* , Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/291482/LIT_8879_df7d5c.pdf

Jones, D. S., Kohl, C., Grettenberger, C., Larson, L. N., Burgos, W. D., & Macalady, J. L. (2015). Geochemical niches of iron-oxidizing acidophiles in acidic coal mine drainage. *Applied and Environmental Microbiology*, 81, 1242-1250.

Kane, S. R., Chakicherla, A. Y., Chain, P. S., Schmidt, R., Shin, M. W., Legler, T. C., ... & Hristova, K. R. (2007). Whole-genome analysis of the methyl *tert*-butyl ether-degrading beta-proteobacterium *Methylibium petroleiphilum* PM1. *Journal of Bacteriology*, *189*(5), 1931-1945.

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28, 27-30.

Kantor, R. S., Zyl, A. W., Hille, R. P., Thomas, B. C., Harrison, S. T., & Banfield, J. F. (2015). Bioreactor microbial ecosystems for thiocyanate and cyanide degradation unravelled with genome-resolved metagenomics. *Environmental Microbiology*, 17, 4929-4941.

Karl, D. M. (2002). Nutrient dynamics in the deep blue sea. *Trends in Microbiology*, 10, 410-418.

Kay, S. E., Clark, R. A., White, K. L., & Peel, M. M. (2001). Recurrent *Achromobacter piechaudii* bacteria in a Patient with Hematological Malignancy. *Journal of Clinical Microbiology*, 39, 808-810.

Kazamia, E., Czesnick, H., Nguyen, T.T., Croft, M.T., Sherwood, E., Sasso, S., Hodson, S.J., Warren, M.J., & Smith, A.G. (2012). Mutualistic interactions

between vitamin B12 -dependent algae and heterotrophic bacteria exhibit regulation. *Environmental Microbiology*, 6, 1462-2920.

Kelly, D. P., & Wood, A. P. (2000). Reclassification of some species of *Thiobacillus* to the newly designated genera *Acidithiobacillus* gen. nov., *Halothiobacillus* gen. nov. and *Thermithiobacillus* gen. nov. *International Journal of Systematic and Evolutionary Microbiology*, 50, 511-516.

Kielak, A. M., Barreto, C. C., Kowalchuk, G. A., van Veen, J. A., & Kuramae, E. E. (2016). The ecology of Acidobacteria: moving beyond genes and genomes. *Frontiers in Microbiology*, *7*.

Kim, B. H., Ramanan, R., Cho, D. H., Oh, H. M., & Kim, H. S. (2014). Role of Rhizobium, a plant growth promoting bacterium, in enhancing algal biomass through mutualistic interaction. *Biomass and Bioenergy*, 69, 95-105.

Kim, J., Kim, M. S., Koh, A. Y., Xie, Y., & Zhan, X. (2016). FMAP: Functional Mapping and Analysis Pipeline for metagenomics and metatranscriptomics studies. *BMC bioinformatics*, 17, 420.

Kovaleva, O. L., Merkel, A. Y., Novikov, A. A., Baslerov, R. V., Toshchakov, S. V., & Bonch-Osmolovskaya, E. A. (2015). *Tepidisphaera mucosa* gen. nov., sp. nov., a moderately thermophilic member of the class Phycisphaerae in the phylum Planctomycetes, and proposal of a new family, Tepidisphaeraceae fam. nov., and a new order, Tepidisphaerales ord. nov. *International Journal of Systematic and Evolutionary Microbiology*, 65, 549-555.

Krueger, F., (2015) Trim Galore. Available: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ [Accessed 22/9/17]

Kumar, P. S., Brooker, M. R., Dowd, S. E., & Camerlengo, T. (2011). Target region selection is a critical determinant of community fingerprints generated by 16S pyrosequencing. *PloS one*, 6, e20956.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, 5, R12.

Kuyukina, M. S., & Ivshina, I. B. (2010). Application of *Rhodococcus* in bioremediation of contaminated environments. *Biology of Rhodococcus*. Springer Berlin Heidelberg, 231-262.

Kuzmanović, N., Prokić, A., Ivanović, M., Zlatković, N., Gašić, K., & Obradović, A. (2015). Genetic diversity of tumorigenic bacteria associated with crown gall disease of raspberry in Serbia. *European Journal of Plant Pathology*, 1-13.

Kyrpides, N. C., Woyke, T., Eisen, J. A., Garrity, G., Lilburn, T. G., Beck, B. J., ... & Klenk, H. P. (2014). Genomic Encyclopedia of Type Strains, Phase I: the one thousand microbial genomes (KMG-I) project. *Standards in Genomic Sciences*, 9, 1278.

Lage, O. M. and Bondoso, J. (2011). Planctomycetes diversity associated with macroalgae. *FEMS Microbiology Ecology*, 78, 366-375.

Lage, O. M., & Bondoso, J. (2012). Bringing Planctomycetes into pure culture. *Frontiers in Microbiology*, 3, DOI: 10.3389/fmicb.2012.00405

Lage, O. M. and Bondoso, J. (2014). Planctomycetes and macroalgae, a striking association. *Terrestrial Microbiology*, 5, 267.

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35*,* 3100-3108.

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., & Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences*, 82, 6955-6959.

Larimer, F. W., Chain, P., Hauser, L., Lamerdin, J., Malfatti, S., Do, L., ... & Tabita, F. R. (2004). Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. *Nature Biotechnology*, 22, 55.

Leal, A. J., Rodrigues, E. M., Leal, P. L., Júlio, A. D. L., Fernandes, R. D. C. R., Borges, A. C., & Tótola, M. R. (2017). Changes in the microbial community during bioremediation of gasoline-contaminated soil. *Brazilian journal of Microbiology*, 48, 342-351.

Lee, A. K., Lewis, D. M., & Ashman, P. J. (2013). Harvesting of marine microalgae by electroflocculation: the energetics, plant design, and economics. *Applied Energy*, *108*, 45-53.

Lee, A. K., Lewis, D. M., & Ashman, P. J. (2009). Microbial flocculation, a potentially low-cost harvesting technique for marine microalgae for the production of biodiesel. *Journal of Applied Phycology*, 21, 559-567.

Lee, J., Cho, D. H., Ramanan, R., Kim, B. H., Oh, H. M., & Kim, H. S. (2013). Microalgae-associated bacteria play a key role in the flocculation of *Chlorella vulgaris*. *Bioresource Technology*, 131, 195-201.

Lee, J., Park, B., Woo, S. G., Lee, J., & Park, J. (2014). *Prosthecobacter algae* sp. nov., isolated from activated sludge using algal metabolites. *International Journal of Systematic and Evolutionary Microbiology*, 64, 663-667.

Lee, M., Woo, S. G., & Ten, L. N. (2011). *Shinella daejeonensis* sp. nov., a nitrate-reducing bacterium isolated from sludge of a leachate treatment plant. *International Journal of Systematic and Evolutionary Microbiology*, 61, 2123-2128.

Lenneman, E. M., Wang, P., & Barney, B. M. (2014). Potential application of algicidal bacteria for improved lipid recovery with specific algae. *FEMS Microbiology letters*, *354*(2), 102-110.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25, 1754-60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 16, 2078-9.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., ... & Yang, B. (2012). Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics*, *11*, 25-37.

Lin, D. X., Wang, E. T., Tang, H., Han, T. X., He, Y. R., Guan, S. H., & Chen, W. X. (2008). *Shinella kummerowiae* sp. nov., a symbiotic bacterium isolated from root nodules of the herbal legume *Kummerowia stipulacea*. *International Journal of Systematic and Evolutionary Microbiology*, 58, 1409-1413.

Lindgreen, S., Adair, K. L., & Gardner, P. P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*, *6*.

List of Prokayotic Names with Standing in Nomenclature (2017) Available: http://www.bacterio.net/index.html [Accessed 23/10/17].

Love, J. (2010). Conference Report: Society for General Microbiology Bioenergy Fuel Sources Symposium. *Biofuels*, 1, 15-17.

Magoc, T., & Salzberg, S. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies, *Bioinformatics,* 27, 2957-63.

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., Tallon, L., & Salzberg, S. L. (2013). GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics*, 29, 1718-1725.

Mann, A. J., Hahnke, R. L., Huang, S., Werner, J., Xing, P., Barbeyron, T., ... & Glöckner, F. O. (2013). The genome of the alga-associated marine *flavobacterium Formosa agariphila* KMM 3901T reveals a broad potential for

degradation of algal polysaccharides. *Applied and Environmental Microbiology*, 79, 6813-6822.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., ... & Dewell, S. B. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437, 376.

Martens, M., Dawyndt, P., Coopman, R., Gillis, M., De Vos, P., & Willems, A. (2008). Advantages of multilocus sequence analysis for taxonomic studies: a case study using 10 housekeeping genes in the genus *Ensifer* (including former *Sinorhizobium*). *International Journal of Systematic and Evolutionary Microbiology*, 58, 200-214.

Martins, M., Faleiro, M. L., Barros, R. J., Veríssimo, A. R., Barreiros, M. A., & Costa, M. C. (2009). Characterization and activity studies of highly heavy metal resistant sulphate-reducing bacteria to be used in acid mine drainage decontamination. *Journal of Hazardous Materials*, 166, 706-713.

Matsui, T., Shinzato, N., Tamaki, H., Muramatsu, M., & Hanada, S. (2009). *Shinella yambaruensis* sp. nov., a 3-methyl-sulfolane-assimilating bacterium isolated from soil. *International Journal of Systematic and Evolutionary Microbiology*, 59, 536-539.

Mattick, J. S. (2002). Type IV pili and twitching motility. *Annual Reviews in Microbiology*, 56, 289-314.

Mayali, X., & Azam, F. (2004). Algicidal bacteria in the sea and their impact on algal blooms. *Journal of Eukaryotic Microbiology*, 51, 139-144.

Mediannikov, O., Sekeyová, Z., Birg, M. L., & Raoult, D. (2010). A novel obligate intracellular gamma-proteobacterium associated with ixodid ticks, *Diplorickettsia massiliensis*, gen. nov., sp. nov. *PloS one*, 5, e11478.

Méndez-García, C., Peláez, A. I., Mesa, V., Sánchez, J., Golyshina, O. V., & Ferrer, M. (2015). Microbial diversity and metabolic networks in acid mine drainage habitats. *Frontiers in Microbiology*, 6, 475.

Metzger, P., & Largeau, C. (2005). *Botryococcus braunii*: a rich source for hydrocarbons and related ether lipids. *Applied Microbiology and Biotechnology*, 66, 486-496.

Meyer, N., Bigalke, A., Kaulfuß, A., & Pohnert, G. (2017). Strategies and ecological roles of algicidal bacteria. *FEMS Microbiology Reviews*, 41, 880-899.

Mielke, R. E., Pace, D. L., Porter, T., & Southam, G. (2003). A critical stage in the formation of acid mine drainage: Colonization of pyrite by *Acidithiobacillus ferrooxidans* under pH-neutral conditions. *Geobiology*, 1, 81-90.

Mignard, S., & Flandrois, J. P. (2006). 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. *Journal of Microbiological Methods*, 67, 574-581.

Minot, S. S., Krumm, N., & Greenfield, N. B. (2015). One codex: a sensitive and accurate data platform for genomic microbial identification. *bioRxiv*, 027607.
Müller, R. H., Rohwerder, T., & Harms, H. (2008). Degradation of fuel oxygenates and their main intermediates by *Aquincola tertiaricarbonis* L108. *Microbiology*, 154, 1414-1421.

Mirete, S., De Figueras, C. G., & González-Pastor, J. E. (2007). Novel nickel resistance genes from the rhizosphere metagenome of plants adapted to acid mine drainage. *Applied and Environmental Microbiology*, 73, 6001-6011.

Monciardini, P., Cavaletti, L., Schumann, P., Rohde, M., & Donadio, S. (2003). *Conexibacter woesei* gen. nov., sp. nov., a novel representative of a deep evolutionary line of descent within the class Actinobacteria. *International Journal of Systematic and Evolutionary Microbiology*, 52, 569-576.

Mukherjee, S., Seshadri, R., Varghese, N. J., Eloe-Fadrosh, E. A., Meier-Kolthoff, J. P., Göker, M., ... & Yoshikuni, Y. (2017). 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nature Biotechnology*, 35, 676.

Narzisi, G., & Mishra, B. (2011). Comparing de novo genome assembly: the long and short of it. *PloS one*, 4*,* e19175.

Natarajan, K. A. (2008). Microbial aspects of acid mine drainage and its bioremediation. *Transactions of Nonferrous Metals Society of China*, 18, 1352-1360.

National Centre for Biotechnology Information (NCBI) (2017). 16S RefSeq records processing and curation. Available: https://www.ncbi.nlm.nih.gov/refseq/targetedloci/16S_process/ [Accessed 22/9/17]

Neal, C., Whitehead, P. G., Jeffery, H., & Neal, M. (2005). The water quality of the River Carnon, west Cornwall, November 1992 to March 1994: the impacts of Wheal Jane discharges. *Science of the Total Environment*, 338, 23-39.

Nedashkovskaya, O. I., Balabanova, L. A., Zhukova, N. V., Kim, S. J., Bakunina, I. Y., & Rhee, S. K. (2014). *Flavobacterium ahnfeltiae* sp. nov., a new marine polysaccharide-degrading bacterium isolated from a Pacific red alga. *Archives of Microbiology*, 196, 745-752.

Nichols, D., Lewis, K., Orjala, J., Mo, S., Ortenberg, R., O'Connor, P., Zhao, C., Vouros, P., Kaeberlein, T. & Epstein, S. (2008) Short peptide induces an "uncultivable" microorganism to grow in vitro. *Applied and Environmental Microbiology*, 74, 4889–4897.

Nilakanta, H., Drews, K. L., Firrell, S., Foulkes, M. A., & Jablonski, K. A. (2014). A review of software for analyzing molecular sequences. *BMC research notes*, 7, 830.

Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulos, G. A., Papanikolaou, N., Kotoulas, G., ... & Iliopoulos, L. (2015). Metagenomics: tools and insights for analysing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and Biology Insights*, *9*, BBI-S12462.

Oggerin, M., Arahal, D. R., Rubio, V., & Marín, I. (2009). Identification of *Beijerinckia fluminensis* strains CIP 106281T and UQM 1685T as *Rhizobium radiobacter* strains, and proposal of *Beijerinckia doebereinerae* sp. nov. to accommodate Beijerinckia fluminensis LMG 2819. *International Journal of Systematic and Evolutionary Microbiology*, 59, 2323-2328.

Olguín, E. J. (2012). Dual purpose microalgae–bacteria-based systems that treat wastewater and produce biodiesel and chemical products within a Biorefinery. *Biotechnology Advances*, 30*,* 1031-1046.

Oliveros, J.C. (2007-2015) Venny. An interactive tool for comparing lists with Venn's diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html

Ozer, E. A., Allen, J. P., & Hauser, A. R. (2014). Characterization of the core and accessory genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt, *BMC Genomics*, 15, 737.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25, 1043-1055.

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P. A., Woodcroft, B. J., Evans, P. N., ... & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, 2, 1533.

Parte, A.C. (2013). LPSN — list of prokaryotic names with standing in nomenclature. *Nucleic Acids Research*, 42, D613–D616; doi: 10.1093/nar/gkt1111

Peabody, M. A., Van Rossum, T., Lo, R., & Brinkman, F. S. (2015). Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities, *BMC Bioinformatics*, 16, 362.

Piddock, L. J. (2006). Multidrug-resistance efflux pumps, not just for resistance. *Nature Reviews Microbiology*, 4, 629-636.

Pinto, U. M., Pappas, K. M., & Winans, S. C. (2012). The ABCs of plasmid replication and segregation. *Nature Reviews Microbiology*, 10, 755-765.

Pirrie, D., Power, M. R., Rollinson, G., Camm, G. S., Hughes, S. H., Butcher, A. R., & Hughes, P. (2003). The spatial distribution and source of arsenic, copper, tin and zinc within the surface sediments of the Fal Estuary, Cornwall, UK. *Sedimentology*, 50, 579-595.

Plummer, E., Twin, J., Bulach, D. M., Garland, S. M., & Tabrizi, S. N. (2015). A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. *Journal of Proteomics & Bioinformatics*, 8*,* 283.

Poehlein, A., Freese, H., Daniel, R., & Simeonova, D. D. (2016). Genome sequence of *Shinella* sp. strain DD12, isolated from homogenized guts of starved *Daphnia magna*. *Standards in Genomic Sciences*, 11, 14.
Prakash, O., Green, S. J., Jasrotia, P., Overholt, W. A., Canion, A., Watson, D. B., ... & Kostka, J. E. (2012). *Rhodanobacter denitrificans* sp. nov., isolated from nitrate-rich zones of a contaminated aquifer. *International journal of Systematic and Evolutionary Microbiology*, 62, 2457-2462.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 21, 7188-7196.

Qin, J. (2005). Bio-hydrocarbons from Algae. Impacts of temperature, light and salinity on algal growth. *Australian Government Rural Industries Research and Development Corporation.* RIRDC Publication No 05/025

Qiu, J., Yang, Y., Zhang, J., Wang, H., Ma, Y., He, J., & Lu, Z. (2016). The Complete Genome Sequence of the Nicotine-Degrading Bacterium *Shinella* sp. HZN7. *Frontiers in Microbiology*, *7*, 1348

Quaiser, A., Ochsenreiter, T., Lanz, C., Schuster, S. C., Treusch, A. H., Eck, J. & Schleper, C. (2003). Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. *Molecular Microbiology*, 50, 563–575.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1), D590-D596.

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., ... & Ouédraogo, N. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530, 228.

Quinlan, A. R., & Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 6, 841-842.

Ramanan, R., Kim, B. H., Cho, D. H., Oh, H. M., & Kim, H. S. (2016). Algae–bacteria interactions: evolution, ecology and emerging applications. *Biotechnology Advances*, 34, 14-29.

Rappé, M. S., & Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Reviews in Microbiology*, 57, 369-394.

Rhoads, A., & Au, K. F. (2015). PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, 13, 278-289.

Richards, T. & Bass, D. (2005) Molecular screening of free-living microbial eukaryotes: Diversity and distribution using a meta-analysis. *Current Opinion in Microbiology*, 5, 240-252.

Ridderberg, W., Wang, M., & Nørskov-Lauritsen, N. (2012). Multilocus sequence analysis of isolates of *Achromobacter* from patients with cystic fibrosis reveals infecting species other than *Achromobacter xylosoxidans*. *Journal of Clinical Microbiology*, 8, 2688-2694.

Rivas, M. O., Vargas, P., & Riquelme, C. E. (2010). Interactions of *Botryococcus braunii* cultures with bacterial biofilms. *Microbial Ecology*, 60, 628-635.

Rodrigue, S., Materna, A., Timberlake, S., Blackburn, M., Malmstrom, R., Alm, E., & Chisholm, S. (2010) Unlocking short read sequencing for metagenomics. *PloS one*, 5, e11840.

Rojas, C., Gutierrez, R. M., & Bruns, M. A. (2016). Bacterial and eukaryal diversity in soils forming from acid mine drainage precipitates under reclaimed vegetation and biological crusts. *Applied Soil Ecology*, *105*, 57-66.

Ronen, Z., Visnovsky, S., & Nejidat, A. (2005). Soil extracts and co-culture assist biodegradation of 2, 4, 6-tribromophenol in culture and soil by an auxotrophic *Achromobacter piechaudii* strain TBPZ. *Soil Biology and Biochemistry*, 37*,* 1640-1647.

Rosen, B. P. (1990). The plasmid-encoded arsenical resistance pump: an anion-translocating ATPase. *Research in Microbiology*, 141, 336-341.

Rowe, O. F., Sánchez‐España, J., Hallberg, K. B., & Johnson, D. B. (2007). Microbial communities and geochemical dynamics in an extremely acidic, metal‐rich stream at an abandoned sulfide mine (Huelva, Spain) underpinned by two functional primary production systems. *Environmental Microbiology*, 9, 1761-1771.

Ruberto, L. A., Vazquez, S., Lobalbo, A., & Mac Cormack, W. P. (2005). Psychrotolerant hydrocarbon-degrading *Rhodococcus* strains isolated from polluted Antarctic soils. *Antarctic Science*, 17, 47-56.

Rubin, E. M. (2008). Genomics of cellulosic biofuels. *Nature*, 454, 841.

Sánchez-Andrea, I., Rodríguez, N., Amils, R., & Sanz, J. L. (2011). Microbial diversity in anaerobic sediments at Rio Tinto, a naturally acidic environment with a high heavy metal content. *Applied and Environmental Microbiology*, 77, 6085-6093.

Sánchez-Andrea, I., Knittel, K., Amann, R., Amils, R., & Sanz, J. L. (2012). Quantification of Tinto River sediment microbial communities: importance of sulfate-reducing bacteria and their role in attenuating acid mine drainage. *Applied and Environmental Microbiology*, 78, 4638-4645.

Sandkvist, M. (2001). Type II secretion and pathogenesis. *Infection and Immunity*, 69, 3523-3535.

Sangwan, N., Xia, F., & Gilbert, J. A. (2016). Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, *4*(1), 8.

Santofimia, E., González-Toril, E., López-Pamo, E., Gomariz, M., Amils, R., & Aguilera, Á. (2013). Microbial diversity and its relationship to physicochemical characteristics of the water in two extreme acidic pit lakes from the Iberian Pyrite Belt (SW Spain). *PLoS One*, 8, e66746.

Santos, P., Pinhal, I., Rainey, F. A., Empadinhas, N., Costa, J., Fields, B., ... & da Costa, M. S. (2003). Gamma-Proteobacteria *Aquicella lusitana* gen. nov., sp. nov., and *Aquicella siphonis* sp. nov. infect protozoa and require activated charcoal for growth in laboratory media. *Applied and Environmental Microbiology*, 69*,* 6533-6540.

Schäfer, F., Breuer, U., Benndorf, D., Von Bergen, M., Harms, H., & Müller, R. H. (2007). Growth of *Aquincola tertiaricarbonis* L108 on tert-Butyl Alcohol Leads to the Induction of a Phthalate Dioxygenase-related Protein and its Associated Oxidoreductase Subunit. *Engineering in Life Sciences*, 7, 512-519.

Schippers, A., Bosecker, K., Spröer, C., & Schumann, P. (2005). *Microbacterium oleivorans* sp. nov. and *Microbacterium hydrocarbonoxydans* sp. nov., novel crude-oil-degrading Gram-positive bacteria. *International Journal of Systematic and Evolutionary Microbiology*, 55, 655-660.

Schlieper, D., Oliva, M. A., Andreu, J. M., & Löwe, J. (2005). Structure of bacterial tubulin BtubA/B: evidence for horizontal gene transfer. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 9170-9175.

Schmidt, T., DeLong, E. & Pace, N. (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *Journal of Bacteriology*, 173, 4371–8.

Schoch, P. E., & Cunha, B. A. (1988). Nosocomial *Achromobacter xylosoxidans* infections. *Infection Control*, 9, 84-87.

Schuster, S. C. (2007) Next-generation sequencing transforms today's biology. *Nature*, 5, 16-18.

Scott, S. A., Davey, M. P., Dennis, J. S., Horst, I., Howe, C. J., Lea-Smith, D. J., & Smith, A. G. (2010). Biodiesel from algae: challenges and prospects. *Current opinion in biotechnology*, 21, 277-286.

Searchinger, T., Heimlich, R., Houghton, R. A., Dong, F., Elobeid, A., Fabiosa, J., ... & Yu, T. H. (2008). Use of US croplands for biofuels increases greenhouse gases through emissions from land-use change. *Science*, 319, 1238-1240.

Sheng, X. F., He, L. Y., Zhou, L., & Shen, Y. Y. (2009). Characterization of *Microbacterium* sp. F10a and its role in polycyclic aromatic hydrocarbon removal in low-temperature soil. *Canadian Journal of Microbiology*, 55, 529-535.

Sheoran, A. S., & Sheoran, V. (2006). Heavy metal removal mechanism of acid mine drainage in wetlands: a critical review. *Minerals Engineering*, 19, 105-116.

Sherpa, R. T., Reese, C. J., & Aliabadi, H. M. (2015). Application of iChip to grow “uncultivable” microorganisms and its impact on antibiotic discovery. *Journal of Pharmacy & Pharmaceutical Sciences*, 18, 303-315.

Simon, C., & Daniel, R. (2010) Metagenomic Analyses: Past and Future Trends. *Applied and Environmental Biotechnology*, 77, 1153-1161.

Singh, J., & Gu, S. (2010). Commercialization potential of microalgae for biofuels production. *Renewable and Sustainable Energy Reviews*,14, 2596-2610.

Skousen, J., Zipper, C. E., Rose, A., Ziemkiewicz, P. F., Nairn, R., McDonald, L. M., & Kleinmann, R. L. (2017). Review of passive systems for acid mine drainage treatment. *Mine Water and the Environment*, 36, 133-153.

Solano-Serena, F., Marchal, R., Casarégola, S., Vasnier, C., Lebeault, J. M., & Vandecasteele, J. P. (2000). A *Mycobacterium* strain with extended capacities for degradation of gasoline hydrocarbons. *Applied and Environmental Microbiology*, 66, 2392-2399.

Staley, J. T., Bont, J. A. M.& Jonge, K. (1976). *Prosthecobacter fusiformis* nov. gen. et sp., the fusiform caulobacter*. Antonie van Leeuwenhoek,* 42, 333–342.

Staley, J. T., & Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Reviews in Microbiology*, 39, 321-346.

Stolze, Y., Zakrzewski, M., Maus, I., Eikmeyer, F., Jaenicke, S., Rottmann, N., ... & Schlüter, A. (2015). Comparative metagenomics of biogas-producing microbial communities from production-scale biogas plants operating under wet or dry fermentation conditions. *Biotechnology for Biofuels*, 8*,*14.

Su, X., Xu, J., & Ning, K. (2012). Parallel-META: efficient metagenomic data analysis based on high-performance computation. *BMC Systems Biology*, 6, S16.

Swale, E. M. F. (1968). The phytoplankton of Oak Mere, Cheshire, 1963–1966. *British Phycological Bulletin*, 3, 441-449.

Tanabe, Y., Kato, S., Matsuura, H., & Watanabe, M. M. (2012). A *Botryococcus* strain with bacterial ectosymbionts grows fast and produces high amount of hydrocarbons. *Procedia Environmental Sciences*, *15*, 22-26.

Tang, Y. Z., Koch, F., & Gobler, C. J. (2010). Most harmful algal bloom species are vitamin B1 and B12 auxotrophs. *Proceedings of the National Academy of Sciences*, 107, 20756-20761.

Tanoi, T., Kawachi, M., & Watanabe, M. M. (2011). Effects of carbon source on growth and morphology of *Botryococcus braunii*. *Journal of Applied Phycology*, 23, 25-33.

Takeda M., Yoneya A., Miyazaki Y., Kondo K., Makita H., Kondoh M., Suzuki I. & Koizumi J.-I. (2008*). Prosthecobacter fluviatilis* sp. nov., which lacks the bacterial tubulin *btubA* and *btubB* genes. *Int J Syst Evol Microbiol,* 58, 1561-1565.

Taylor, M., Mediannikov, O., Raoult, D., & Greub, G. (2012). Endosymbiotic bacteria associated with nematodes, ticks and amoebae. *FEMS Immunology & Medical Microbiology*, 64, 21-31.

Thomas, M. C., Thomas, D. K., Selinger, L. B., & Inglis, G. D. (2011). SPYDER, a new method for in silico design and assessment of 16S rRNA gene primers for molecular microbial ecology. *FEMS Microbiology Letters*, 320, 152-159.

Tomich, M., Planet, P. J., & Figurski, D. H. (2007). The tad locus: postcards from the widespread colonization island. *Nature Reviews Microbiology*, 5, 363-375.

Trumm, D. (2010). Selection of active and passive treatment systems for AMD—flow charts for New Zealand conditions. *New Zealand Journal of Geology and Geophysics*, 53, 195-210.

Tseng, T. T., Tyler, B. M., & Setubal, J. C. (2009). Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. *BMC Microbiology*, 9 (Suppl 1), S2.

Tyson, G.W., Chapman, J., & Hugenholtz, P. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428, 37–43.

Tyson, G. W., Lo, I., Baker, B. J., Allen, E. E., Hugenholtz, P., & Banfield, J. F. (2005). Genome-directed isolation of the key nitrogen fixer Leptospirillum ferrodiazotrophum sp. nov. from an acidophilic microbial community. *Applied and Environmental Microbiology*, 71, 6319-6324.

Uduman, N., Qi, Y., Danquah, M. K., Forde, G. M., & Hoadley, A. (2010). Dewatering of microalgal cultures: a major bottleneck to algae-based fuels. *Journal of Renewable and Sustainable Energy*, 2, 701.

Ummalyma, S. B., Gnansounou, E., Sukumaran, R. K., Sindhu, R., Pandey, A., & Sahoo, D. (2017). Bioflocculation: An alternative strategy for harvesting of microalgae–An overview. *Bioresource Technology*, *242*, 227-235.

Utturkar, S. M., Klingeman, D. M., Land, M. L., Schadt, C. W., Doktycz, M. J., Pelletier, D. A., & Brown, S. D. (2014). Evaluation and validation of *de novo* and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics*, 30, 2709-2716.

Valdés, J., Pedroso, I., Quatrini, R., Dodson, R. J., Tettelin, H., Blake, R., ... & Holmes, D. S. (2008). *Acidithiobacillus ferrooxidans* metabolism: from genome sequence to industrial applications. *BMC Genomics*, 9, 597.

Van de Peer, Y., Chapelle, S., & De Wachter, R. (1996). A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic Acids Research*, 24, 3381-3391.

Van Veen, E. M., Lottermoser, B. G., Parbhakar-Fox, A., Fox, N., & Hunt, J. (2016). A new test for plant bioaccessibility in sulphidic wastes and soils: A case study from the Wheal Maid historic tailings repository in Cornwall, UK. *Science of the Total Environment*, 563, 835-844.

Vartoukian, S., Palmer, R. & Wade, W. (2010) Strategies for culture of "unculturable" bacteria. *FEMS Microbiology Letters*, 309, 1–7.
Vaz-Moreira, I., Faria, C., Lopes, A. R., Svensson, L. A., Moore, E. R., Nunes, O. C., & Manaia, C. M. (2010). *Shinella fusca* sp. nov., isolated from domestic waste compost. *International Journal of Systematic and Evolutionary Microbiology*, 60, 144-148.

Velázquez, E., Peix, A., Zurdo-Piñiro, J. L., Palomo, J. L., Mateos, P. F., Rivas, R., ... & Martínez-Molina, E. (2005). The coexistence of symbiosis and pathogenicity-determining genes in *Rhizobium rhizogenes* strains enables them to induce nodules and tumors or hairy roots in plants. *Molecular Plant-Microbe Interactions*, 18, 1325-1332.

Venter, J.C., Remington, K., & Heidelberg, J.F. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304, 66–74.

Vera, M., Schippers, A., & Sand, W. (2013). Progress in bioleaching: fundamentals and mechanisms of bacterial metal sulfide oxidation—part A. *Applied Microbiology and Biotechnology*, 97, 7529-7541.

Vogan, A. A., & Higgs, P. G. (2011). The advantages and disadvantages of horizontal gene transfer and the emergence of the first species. *Biol Direct*, 6.

Wallden, K., Rivera-Calzada, A., & Waksman, G. (2010). Microreview: Type IV secretion systems: versatility and diversity in function. *Cellular Microbiology*, 12, 1203-1212.

Waltz, E. (2013) Algal biofuels questioned. *Nature Biotechnology*, 31, 12.

Wan, C., Zhao, X. Q., Guo, S. L., Alam, M. A., & Bai, F. W. (2013). Bioflocculant production from Solibacillus silvestris W01 and its application in cost-effective harvest of marine microalga Nannochloropsis oceanica by flocculation. *Bioresource Technology*, 135, 207-212.

Wang, H., Hill, R. T., Zheng, T., Hu, X., & Wang, B. (2016). Effects of bacterial communities on biofuel-producing microalgae: stimulation, inhibition and harvesting. *Critical Reviews in Biotechnology*, 36, 341-352.

Wang, Z. Gerstein, M., & Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews*, 10, 57-63.

Whitehead, P. G., Cosby, B. J., & Prior, H. (2005). The Wheal Jane wetlands model for bioremediation of acid mine drainage. *Science of the Total Environment*, 338, 125-135.

Whitehead, P. G., & Prior, H. (2005). Bioremediation of acid mine drainage: an introduction to the Wheal Jane wetlands project. *Science of the Total Environment*, *338*, 15-21

Whitman, W.B., Coleman, D.C. & Wiebe, W.J. (1998) Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, 95, 6578–6583.

Willems, A., & Collins, M. D. (1993). Phylogenetic analysis of *Rhizobia* and *Agrobacteria* based on 16S rRNA gene sequences. *International Journal of Systematic Bacteriology*, 43, 305-313.

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences,* 74, 5088-5090.

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15, R46.

Wolf, F. (1983). *Botryococcus braunii* an unusual hydrocarbon-producing alga. *Applied Biochemistry and Biotechnology*. 8, 249-260.

Xiao, C. L., Chen, Y., Xie, S. Q., Chen, K. N., Wang, Y., Han, Y., ... & Xie, Z. (2017). MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *nature methods*, 14, 1072.

Xing, M. N., Zhang, X. Z., & Huang, H. (2012). Application of metagenomic techniques in mining enzymes from microbial communities for biofuel synthesis. *Biotechnology Advances*, 30, 920-929.

Yamaguchi, K., Nakano, H., Murakami, M., Konosu, S., Nakayama, O., Kanda, M., Nakamura, A. & Iwamoto, H. (1987). Lipid composition of a green alga, *Botryococcus* braunii. *Agricultural and biological chemistry*, 51, 493-498.

Yelverton, E., Leung, D., Week, P., Gray, P. W., & Goeddel, D. V. (1981). Bacterial synthesis of a novel human leukocyte interferon. *Nucleic Acids Research*, 9, 731-741.

Yoon, J., Jang, J. H., & Kasai, H. (2014). *Algisphaera agarilytica* gen. nov., sp. nov., a novel representative of the class Phycisphaerae within the phylum Planctomycetes isolated from a marine alga. *Antonie van Leeuwenhoek*, 105, 317-324.

Young, J. M., Kuykendall, L. D., Martinez-Romero, E., Kerr, A., & Sawada, H. (2001). A revision of Rhizobium Frank 1889, with an emended description of the genus, and the inclusion of all species of *Agrobacterium Conn* 1942 and *Allorhizobium undicola* de Lajudie et al. 1998 as new combinations: *Rhizobium radiobacter, R. rhizogenes, R. rubi, R. undicola* and *R. vitis*. *International Journal of Systematic and Evolutionary Microbiology*, 51, 89-103.

Younger, P. L., Coulton, R. H., & Froggatt, E. C. (2005). The contribution of science to risk-based decision-making: lessons from the development of full-scale treatment measures for acidic mine waters at Wheal Jane, UK. *Science of the Total Environment*, 338, 137-154.

Zammit, C. M., Mangold, S., Rao Jonna, V., Mutch, L. A., Watling, H. R., Dopson, M., & Watkin, E. L. (2012). Bioleaching in brackish waters—effect of chloride ions on the acidophile population and proteomes of model species. *Applied Microbiology and Biotechnology*, 93, 319-329.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*, 18*,* 821-829

Ziegler, S., Waidner, B., Itoh, T., Schumann, P., Spring, S., & Gescher, J. (2013). *Metallibacterium scheffleri* gen. nov., sp. nov., an alkalinizing gammaproteobacterium isolated from an acidic biofilm. *International journal of Systematic and Evolutionary Microbiology*, 63, 1499-1504.